COMPUTATIONAL TOOLS FOR MOLECULAR EPIDEMIOLOGY AND COMPUTATIONAL GENOMICS OF *NEISSERIA MENINGITIDIS*

A Dissertation
Presented to
The Academic Faculty

By

Lee Scott Katz

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics

Georgia Institute of Technology

December, 2010

COMPUTATIONAL TOOLS FOR MOLECULAR EPIDEMIOLOGY AND COMPUTATIONAL GENOMICS OF *NEISSERIA MENINGITIDIS*

Approved by:

Dr. I. King Jordan, Advisor
School of Biology
Georgia Institute of Technology

Dr. Leonard W. Mayer, Co-advisor
Meningitis Laboratory
Centers for Disease Control and Prevention

Dr. Jung Choi
School of Biology
Georgia Institute of Technology

Dr. Nicholas H. Bergman
School of Biology
Georgia Institute of Technology

Dr. Joshua S. Weitz
School of Biology
Georgia Institute of Technology

Date Approved: October 28, 2010

At the end of the 1990s it was apparent that a universal meningococcal vaccine was beyond the reach of conventional vaccinology…Therefore, the complete genome sequence of [*Neisseria meningitidis*] was determined…

Serruto and Rappuoli, 2006

To my family and friends, for always supporting me…

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor I. King Jordan. He has given me so much support in every aspect of my work for my Ph.D. degree. I have learned everything about the academic world from him. He is an inspiration for my future work. I would not have such a great topic without the world-class scientist Dr. Leonard Mayer. Leonard is an expert on *Neisseria meningitidis*, which makes me really fortunate to have him as a co-advisor. Leonard and King have been grooming me since day one to be a bioinformatician in the Meningitis and Vaccine Preventable Diseases Branch at CDC, for which I am very grateful. I would like to thank my other committee members Drs. Nick Bergman, Joshua Weitz, and Jung Choi. These three professors brought their own insights into my thesis to make it great.

I have been in other research labs before which added to my background. My research career began when I was an undergraduate at Emory in a *Bacillus subtilis* laboratory with Dr. Charles Moran and Dr. Amrita Kumar. In George Wilmont's laboratory at Emory the following year, I gained more insight into wet laboratory practices but realized that the wet lab was not for me. At Georgia Tech in my masters program, Dr. Soojin Yi was my first bioinformatics P.I. and taught me what it was to be in a bioinformatics lab. Dr. Navin Elango, then a graduate student in her lab at the time, was the perfect mentor. I continued from the masters program into the Ph.D. program and entered into Dr. J. Todd Streelman's cichlid laboratory and became the first bioinformatician there. I appreciate being given the opportunity to be a bioinformatics pioneer in a pure wet lab. It made me understand the dynamics between bench scientists and bioinformaticians. My fifth P.I. King was ultimately my best fit and I knew I was in the right lab when I joined. I could not imagine having switched away from his lab once I joined.

I would like to thank everyone in King's lab over the years. In the course of my projects I have had several team members work with me which consisted of Ph.D., M.S., and undergraduate students: Andrey Kislyuk, Nitya Sharma, Andrew Conley, Sonia Agrawal, Pushkala Jayaraman, Viswateja Nelakuditi, Jay Humphrey, Sandeep Namburi, Rob Taylor, and Chris Bolen. Jittima (Jing) Piriyapongsa and Ahsan Huda were my lab mates for some years, and it was inevitable that we became such close friends. Dissertations such as Jing's, Ahsan's, and my own are bittersweet because it also means parting with them, my close friends. Additionally Jing and Ahsan have helped me considerably over the years on a professional level. Others in my lab who have helped me in some way include Jianrong Wang, Daudi Jjingo, Eishita Tyagi, Mark Rutledge, Aditya Pai, Karthik Kota, Madhumati Gundapuneni, and Eddie Loh. Each has given his or her own insights and strengths into my work to make it as great as it was, and each has or will move onto great things.

I would also like to thank the Compgenomics class for the years 2008-2010. These students used real-world data and problems to come up with innovative solutions. I was happy to be the TA for this class and at the same time learn from them. In fact, I would like to give thanks to the future Compgenomics classes which will undoubtedly contribute further to my efforts in public health.

The Meningitis Laboratory and even the Meningitis and Vaccine Preventable Diseases Branch at CDC has been extremely helpful and supportive of my thesis over the years. Although there are many who have been influential to me including Leonard Mayer, I would like to specifically thank Brian Harcourt, Xin Wang, Jennifer Dolan, and Raydel Mair who have helped me significantly. From The Sequencing Activity Lab in the Influenza Division at CDC, I would like to thank John Barnes, James Smagala, and Sheila Bashirian for their help with understanding influenza.

My friends have been very supportive over the years. I am fortunate to have known James Michelich since I was 4 years old, and I know that I would not be as successful today as I am now had he not been there for me. Catherine Carlson has also contributed some sanity while I wrote my thesis. I would like to thank my friend and classmate Taylor Updegrove with whom I now have many great experiences and memories. I would like to thank Elizabeth and Eric Lynch who both helped me adapt to Georgia Tech and who I have known since the beginning of high school. My thanks go out to my friends Laura, Emily, Annie, Karen, Betsy, and all the others who I collectively call LEAK. You guys managed to get me out of the house more often than you realized and helped me balance work and play.

I would like to thank Wenhan Zhu who was my TA in the core bioinformatics course. Wenhan continues to aid me whenever I seek help and is thankfully very selfless. Eddie (Yong-Hwee) Loh has been my classmate, labmate, and friend since the beginning of the Ph.D. program and I am thankful to know him. Darlene Wagner has been in the graduate program with me since the masters program and it is great to have such a colleague with me along the way.

I have no words to express how great my girlfriend Samantha Stuckey has been during the time I have been writing my thesis. She has given me unending love and support and has always lent an ear when I need to discuss anything.

My parents—Drs. Ian and Sheri Katz—are the best parents anyone could have and have always given me the best guidance. My siblings—Jeffrey Katz and Rebecca Katz—are sources of pride for me, and I think about them every day and how great they are. I can sincerely say that I would be nowhere near where I am today without my family.

To anyone I might have forgotten, please accept my apology and know that I am still thankful. I stand on the shoulders of giants, and this thesis is an acknowledgement and a testament to that fact.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

API	Application programming interface
CC	Clonal complex
CDC	Centers for Disease Control and Prevention
CG-Pipeline	Computation genomics pipeline
DUS	DNA uptake sequence
ET	Electrophretic type
fHBP	Factor H-binding protein
GUI	Graphical user interface
HA	Hemagglutinin
InGen	Influenza genomics platform
LOS	Lipooligosaccharide
LPS	Lipopolysaccharide
MGIP	Meningococcus Genome Informatics Platform
MLST	Multilocus sequence typing
MLEE	Multilocus enzyme electrophoresis
NA	Neuraminidase
NBase	*Neisseria* Base
NG	Nongroupable
SASG	Slide agglutination serogrouping
SG	Serogroup
SNP	Single nucleotide polymorphism

ST	Sequence type
STARS	Sequence Typing Analysis and Retrieval System
STaRS	Small target reassortant screen

SELECT DEFINITIONS

Application programming interface	An interface to a website that enables other software or scripts to interact with it
Clonal complex	
CG-Pipeline	A program that performs genome assembly, feature prediction, and feature annotation
Discriminating SNP	A polymorphic site when comparing two groups of genomes such that it can be used as a marker to distinguish one group of genomes from the other
DNA uptake sequence	A 10 nt sequence that aids in natural neisserial transformation
Influenza Genomics Platform	A suite of tools for the analysis of STaRS data
Meningococcus Genome Informatics Platform	A suite of tools for the analysis of MLST data
***Neisseria* Base**	An online comparative genome database and browser for *Neisseria* genomes
Nongroupability	A state where a meningococcal isolate cannot be attributed to any single serogroup
Slide agglutination serogrouping assay	An antibody-based test to determine the serogroup of an isolate
Sequence type	A profile given to an isolate based on MLST
SNP gene	A gene containing or within 1 kb of a discriminating SNP

Single nucleotide polymorphism	A single base difference when comparing multiple genomes
Sequence Typing Analysis and Retrieval System	A program that analyzes multilocus sequence typing data.
Small target reassortant screen	A method of sequencing the segments in an influenza genome to perform surveillance

SUMMARY

Neisseria meningitidis, also known as the meningococcus, is a gram negative diplococcus that inhabits the human nasopharynx (Rosenstein *et al.*, 2001). *N. meningitidis* is arguably the worldwide leading cause of bacterial meningitis which is a sudden and life threatening disease. Even with proper treatment 10% of all cases are fatal, with an additional 10% sequelae.

Since the early 1900s physicians have been attempting to design antisera and vaccines to combat the meningococcus (Bilukha & Rosenstein, 2005; Flexner, 1913; Goldschneider *et al.*, 1969a; Goldschneider *et al.*, 1969b; Gotschlich *et al.*, 1969a; Gotschlich *et al.*, 1969b; Gotschlich *et al.*, 1969c). *N. meningitidis* can also be combated with antibiotics, but many powerful drugs are meeting resistance (e.g.,Wu *et al.*, 2009).

N. meningitidis is surveyed worldwide using many profiling systems. One of the most widely used includes serogrouping assays, which determine the type of polysaccharide capsule enclosing the isolate (Mothershed *et al.*, 2004; Popovic & Ajello, 2003). However to better study the population structure of *N. meningitidis*, scientists use a system called multilocus sequence typing (MLST). MLST is a molecular profiling system that uses seven conserved loci. These loci are sequenced, and their allelic variants are determined as allele calls. The specific combination of allele calls yields the sequence type (ST) which is the profile of the isolate. STs can be compared and tracked to survey isolates worldwide. Fortunately there is a database of STs called PubMLST, with which a scientist can use to compare their sequences with and determine their own ST or declare novel alleles and STs (Jolley *et al.*, 2004). However unfortunately until 2009, the bioinformatics package available for meningococcal reference labs was not user friendly, was not maintained by the original developers, and was out of date (Chan & Ventress, 2001; Katz *et al.*, 2009).

In chapter 2 of this thesis, I describe how I created a suite of tools called the Meningococcus Genome Informatics Platform (MGIP) to analyze MLST data. MGIP is about five times faster and is more flexible than its predecessor, is user friendly, and is becoming the gold standard in MLST analysis. MGIP has had several advances since its release in 2009 including the expansion to other loci and other organisms. In fact as detailed in chapter 3, the CDC Influenza Division has adapted MGIP into their own typing tool called the Influenza Genomics Platform (InGen). InGen will be used to survey influenza worldwide. State health laboratories will be given InGen in exchange for real-time data from across the nation. InGen has been demonstrated to the Egyptian government and has been the subject at a recent talk at an international influenza meeting in China. MGIP has been accepted by CDC as its main analysis tool for MLST. As such the way has been paved for worldwide acceptance of MGIP and InGen for meningococcal and influenza surveillance.

To better understand *N. meningitidis* a whole genome approach has been adopted at CDC. There are many lingering questions that are best suited for whole genome analysis. For example, only some strains of *N. meningitidis* cause disease whereas the great majority of all strains are not invasive. The trigger that causes strains to switch from virulent to carriage or *vice versa* remains a subject of exploration. Another question that could be answered by genomics pertains to the meningococcal capsule. Most often the capsule type can be identified, but what causes it to become nongroupable, *i.e.* unable to be classified into exactly one capsule type? Whole genome analyses can address this question.

Before these questions can be answered whole genomes must be able to be sequenced and characterized. Since 2005, it has been possible to sequence whole genomes in less than a day (Margulies *et al.*, 2005). However, to correctly assemble and annotate the entire genome is a computational challenge. To this end, I have taken a lead role in an enormous effort to create a

computational genomics pipeline (CG-Pipeline). CG-Pipeline takes 454 pyrosequencing data files as input, assembles sequencing reads, predicts features (*e.g.* coding sequences), and annotates them. Since its inception in June 2010, several prominent institutions have begun using it including CDC, Georgia Institute of Technology, Pacific Biosciences, and the National Biodefenses and Countermeasures Center (NBACC) under the Department of Defense. In addition it is quite possible that other major institutions have begun using CG-Pipeline but have understandably not discussed their projects with our team. CG-Pipeline is discussed in chapters 4 and 5.

The output of CG-Pipeline is a well-annotated but plaintext GenBank file. To properly visualize, characterize, and compare each of these genomes, we developed an online comparative database called *Neisseria* Base (NBase). NBase is the fitting second half of CG-Pipeline in that it is a platform suitable for the storage and analysis of all meningococcal genomes. Currently NBase is capable of comparing multiple genomes to yield differences between them and allows a layman scientist to navigate through any of nearly 30 meningococcal genomes hosted on the database. NBase is described in chapter 5.

With the combination of CG-Pipeline and NBase, it became feasible to study meningococcal genomes by sequencing and characterizing them and to ask certain questions. In chapter 6, I questioned why some meningococcal isolates are virulent while the rest are asymptomatically carried. This has been an open question for decades and has only been addressed on a genomic level in the last eight years with little success (Hotopp *et al.*, 2006; Perrin *et al.*, 2002; Schoen *et al.*, 2008; Snyder & Saunders, 2006; Stabler & Hinds, 2006; Stabler *et al.*, 2005). To address this question, I compared eight virulent genomes against three carriage. In the results, I was able to uncover 801 single nucleotide polymorphisms (SNPs) in the *N. meningitidis* genome that distinguish virulent from carriage genomes. Furthermore, I was able to uncover 113 genes that are associated with the 801 SNPs. By investigating the literature, I was able to filter these 113

genes to ten genes that are involved in virulence. These genes have been offered to the scientific community for further investigation. These SNPs and genes represent the only unambiguous genomic distinction between virulent and carried isolates of *N. meningitidis* known at this time.

Another question that I was able to address was what makes some isolates nongroupable? In chapter 7, I compared a nongroupable isolate against reference genomes to speculate on a genomic level why its phenotype is nongroupable. To this end, I developed two methodologies. One, I mapped all 454 pyrosequencing reads of this isolate to a reference genome to develop a profile. This profile is much more comprehensive than any other current genetic profile and allowed me to determine which reference genome was most related to my own query genome. The second method was to ascertain exact recombination points. The genomic basis of this isolate's nongroupability is most likely due to recombination, and so I found the exact places of recombination. The gene most likely responsible for nongroupability was found to be part of a recombination event and is elaborated upon in chapter 7. This study would not have been possible without whole genome sequences due to the scope of the genomic comparisons involved.

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

HISTORY OF *NEISSERIA MENINGITIDIS*, THE MENINGOCOCCUS

The meningococcus is a gram-negative encapsulated diplococcus that lives in the human nasopharynx (Rosenstein *et al.*, 2001). Sometimes it causes acute invasive disease but usually it is carried as a harmless commensal.

The origin of *N. meningitidis* seems to be very recent. Most believe that the disease, and by extension the meningococcus, first made its appearance in 1805 in France and 1806 in the United States (Danielson & Mann, 1806; Vieusseux, 1806). The physician Anton Weichselbaum first isolated the bacterium in 1887 and dubbed it "*epidemic cerebrospinal meningitidis*" (Weichselbaum, 1887). After it was isolated, advances came more often in the late 19th century and early 20th century such as the ability to differentiate it from *N. gonorrhoeae*, identifying its commensal state, its isolation from the human throat, and even the first antisera (Cartwright, 2006; de Souza & Seguro, 2008). Indeed, the mortality rate in the early 20th century went from about 70% to 25% as a result of the antisera (Flexner, 1913).

Development of a vaccine occurred in the early 20th century where heat-killed meningococci were delivered to a group of American medical students, nurses, and their close contacts (Cartwright, 2006). Next, 4700 US soldiers were given the vaccine and it was shown that the number of disease cases was significantly different between immunized and unimmunized soldiers.

Starting in the 1960s, vaccine development began in the US to produce vaccines against the polysaccharide capsule found in serogroups A and C, which accounted for most meningococcal

disease cases in the world (Goldschneider *et al.*, 1969a; Goldschneider *et al.*, 1969b; Gotschlich *et al.*, 1969a; Gotschlich *et al.*, 1969b; Gotschlich *et al.*, 1969c). It would not be until 2000 that a different technique called reverse vaccinology would be applied to find a vaccine against serogroup B (Rappuoli, 2000), and it would not be until 2008 that a vaccine would be in clinical trials under Novartis's auspices (Rinaudo *et al.*, 2009).

DISEASE

N. meningitidis is named as such because it causes meningitis, the inflammation of the meninges. Meningococcal disease is characterized by purulent meningitis, with a sudden onset of headache, fever, and stiffness of neck (Rosenstein *et al.*, 2001). Occasionally, meningococci can reside in the bloodstream causing meningococcal sepsis, which is also known as meningococcemia. *N. meningitidis*, especially serogroup Y, can also occasionally cause pneumonia or other respiratory problems (Racoosin *et al.*, 1998).

Of those who get meningococcal disease, about 10% die, and of those, about 15% develop long-term sequelae such as deafness, cognitive impairment, and other central nervous system complications (Palmgren, 2009). The current recommendation for those at risk for meningococcal disease is to give regiments of the antibiotics rifampin, ciprofloxacin, and ceftriaxone (Bilukha & Rosenstein, 2005). For those who need immediate treatment, ciprofloxacin is recommended but there are signs of increasing resistance (Wu *et al.*, 2009).

Ultimately the most effective way to address meningococcal disease is to confer immunity using vaccines. Current vaccines target the polysaccharide capsule, which ironically protects the bacterium from opsinophagocytosis by the immune system (Campagne *et al.*, 2000; Miller *et al.*, 2001; Zombre *et al.*, 2007) . Current vaccines can target the polysaccharide capsule of serogroups A, C, W135, and Y. However, the capsule from serogroup B cannot be targeted because the immune system cannot distinguish its polysaccharide from polysialic acid units

found in the human brain (Finne *et al.*, 1983). Therefore a conventional vaccine against the polysaccharide capsule of serogroup B would either be ineffective or would induce autoimmunity.

POPULATION STRUCTURE

The population structure of *N. meningitidis* can be described in several ways. The earliest methods of typing involved using antisera to classify them, which evolved into the aforementioned serogroup classification. The known serogroups are A, B, C, D, 29E (Z'), H, I, K, L, W135, X, Y, and Z (Ashton *et al.*, 1983). However, many experimentalists are leaning to believe that serogroup D exists merely as an extension of serogroup C antisera and therefore there are 12, not 13, known serogroups (Popovic & Ajello, 2003; personal communication, Jennifer Dolan Thomas). The serogroups that are known to be invasive are A, B, C, W135, X, and Y. Of those that are invasive, serogroups B and C predominate in the industrialized world (Figure 1.1). Serogroup A prevails in Africa and Asia. Serogrouping is very fundamental and is in wide use today.

With newer technologies in protein and DNA analysis, typing methods became better in resolution. One notable method of typing, called multilocus enzyme electrophoresis (MLEE), screens for certain cytoplasmic enzymes and determines its electromorphic profile through a gel (Caugant *et al.*, 1986). Each different combination of electromorphs yields an electrophoretic type (ET). MLEE has been mostly abandoned however for a new technique called multilocus sequence typing (MLST) (Maiden *et al.*, 1998). Using MLST, a scientist sequences the DNA of seven loci of housekeeping genes, give or take a few loci depending on the species (Table 1.1). The sequence of each locus yields an allele, and the combination of all alleles determines the isolate's sequence type (ST), which is its profile. In MLST, individual STs are assigned to clonal complexes (CCs), which are closely related clusters of STs, typically with at least 4 allele calls in

common. MLST has many advantages over MLEE: all ETs can be directly translated to STs; data used for MLST can be transferred electronically; the resolution is better due to digital data rather than using a gel and also due to the redundancy of the genetic code (nucleotide sequences are more discriminating than amino acid sequences); and its results are easily comparable in that it is a standardized typing scheme whereas the proteins used in MLEE varied between laboratories. One might argue that using housekeeping genes to detect population structure would not yield the desired resolution; however, *Neisseria* undergoes so many recombination events that housekeeping genes are the best way to describe the population structure in lieu of entire genomes (Holmes *et al.*, 1999; Jolley *et al.*, 2005). For all of these reasons MLST is widely used and is virtually a complete successor of MLEE.

Table 1.1. The seven loci used in MLST for *N. meningitidis.*

Locus	Description
abcZ	putative ABC transporter
adk	adenylate kinase
aroE	shikimate dehydrogenase
fumC	fumarate hydratase
gdh	glucose-6-phosphate dehydrogenase
pdhC	pyruvate dehydrogenase subunit
pgm	phosphoglucomutase

Comparative analysis of clonal complexes uncovered by MLST helps to describe the phylogenetic origins of each meningococcal isolate and also lends insight into trends. One trend for example is that nearly all virulence caused by *N. meningitidis* is caused by relatively few CCs. Another trend is that each CC can be correlated with a serogroup (Table 1.2).

Table 1.2. Common virulent STs and their serogroups, adapted from (Yazdankhah et al., 2004). Major serogroup/CC correlation percentages have been bolded.

		Serogroup					
Clonal Complex	Disease Association odds ratio[a]	A	B	C	Y	W135	NG[b]
ST-11	52 (20-135)	0%	11%	81%	0%	6%	3%
ST-23	0.2 (0.1-0.7)	0%	10%	0%	56%	3%	31%
ST-32	0.9 (0.4-2.2)	0%	82%	6%	3%	0%	10%
ST-35	0.3 (0.1-1.1)	4%	50%	0%	4%	0%	42%
ST-162	0.8 (0.4-1.18)	6%	82%	0%	0%	6%	6%
ST-269	6.1 (0.5-12.8)	0%	86%	0%	5%	0%	9%
ST-41/44	1.1 (0.5-2.3)	0%	78%	2%	2%	0%	17%

[a] 95% confidence interval shown in parentheses

[b] Nongroupable

Most epidemiological studies focus around MLST (Didelot *et al.*, 2009; Holmes *et al.*, 1999; Maiden *et al.*, 1998). However, some protein-based typing systems are still being used, especially subtyping. In meningococcal subtyping, PorA, PorB, and FetA are assessed for more resolution of the population structure. PorA and PorB are porins, which allow small molecules to pass across the outer membrane. FetA (previously, FrpB) is a siderophore that is expressed when iron concentrations are low (Beucher & Sparling, 1995). Variants on the PorB protein sequence (and more recently, the DNA sequence) are used to identify the serotype. Variants on the PorA sequence yield a serosubtype. Standard subtyping notation is given by serogroup: PorA: FetA: ST (CC) (Jolley *et al.*, 2007). One example of a profile designation would be **B:P1.19,15:F5-1:ST-33(cc32)** for serogroup B, PorA type P1.19, FetA type F5-1, and ST-33 with ST-33 belonging to clonal complex ST-32. Because the PorB sequence is more time-consuming, it is recommended that it be only used when more resolution is required.

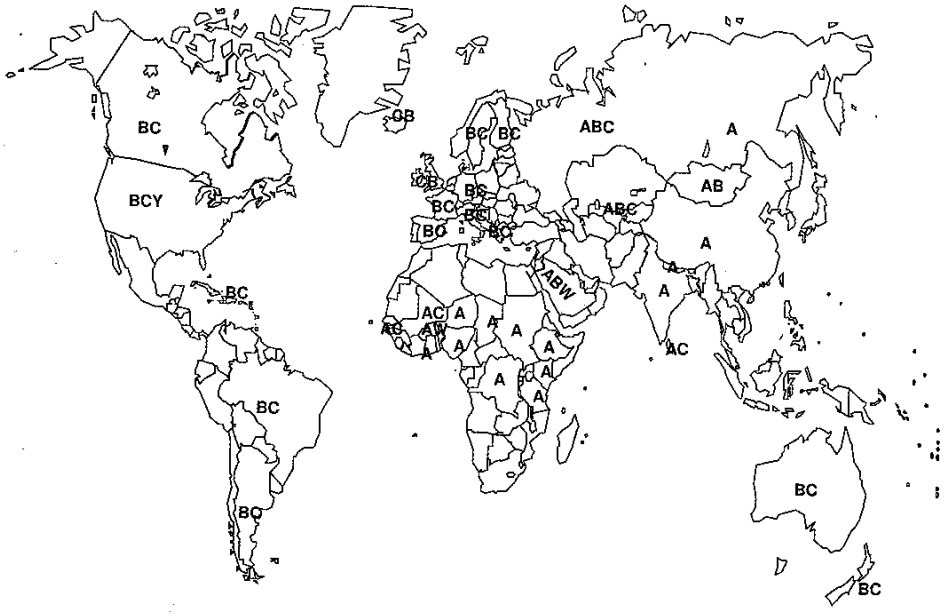


Figure 1.1. Global distribution of meningococcal disease. The distribution of the various serogroups of meningococci is indicated by the appropriate letters. Adapted from (Tikhomirov et al., 1997).

INVASIVE OR CARRIAGE?

Most often *N. meningitidis* can be found asymptomatically carried in the back of the throat (Claus *et al.*, 2005; Dolan-Livengood *et al.*, 2003). However it has been noted that there are a few hypervirulent lineages that are found to cause disease. Many studies have tried to find the underlying genetic cause (Bille *et al.*, 2005; Hotopp *et al.*, 2006; Perrin *et al.*, 2002; Schoen *et al.*, 2008; Snyder & Saunders, 2006; Stabler & Hinds, 2006; Stabler *et al.*, 2005). However, the best results from these studies have been shown support against them, and it is still an open-ended question. There are many hypotheses as to which factor is the genetic switch from invasive to carriage. For example, many believe the answer to be related to pili (Nassif *et al.*, 1997; Stephens *et al.*, 1983), and others believe it the switch is related to evading the immune system with proteins such as IgA protease (Lin *et al.*, 1997; Mistry & Stockley, 2006). While the genetic switch is still unknown, the search is narrowing down the possibilities.

MOLECULAR ASPECTS OF MENINGOCOCCUS

DNA

Sequencing of the meningococcal genome began with the strains Z2491 (Parkhill *et al.*, 2000) and MC58 (Tettelin *et al.*, 2000). Other genomes have been published since then (Table 1.3). In my collaboration with the CDC Meningitis Laboratory we have sequenced several genomes and several are still in the process of being made public (Table 1.3). Some basic facts have arisen as a result of these sequencing projects. The meningococcal genome is about 2.15 Mb; the GC content is, on average, higher than 51%; and the number of genes on average per genome is more than 2100 (unpublished data).

Table 1.3. Completed known genomes at the time of this dissertation.

ID	Geographic Origin [a]	Year isolated [a]	Sg [b]	Disease	ST [c]	CC [c]	PMID [d]
M20918	Iowa, USA	2009	A	Invasive	4789	5	N/A
M13220	Philippines	2005	A	Invasive	7	5	20519285
M18575	Burkina Faso	2003	A	Invasive	2859	5	20519285
Z2491	Gambia	1983	A	Invasive	4	4	10761919
M17277	Maryland, USA	2006	NG	Carriage	5916	41/44	N/A
M16207	North Dakota, USA	2007	B	Invasive	162	162	N/A
M17094	Minnesota, USA	2008	B (C)	Carriage	32	32	N/A
MC58	Gloucester, UK	1985	B	Invasive	74	32	10710307
M10699	Oregon, USA	2003	B	Invasive	32	32	20519285
M5178	Oregon, USA	1998	B	Invasive	32	32	20519285
053442	Anhui province, China	2003	C	Invasive	4821	4821	18031983
8013	France	1989	C	Invasive	177	18	19818133
FAM18	North Carolina, USA	1980s	C	Invasive	11	11	17305430
M15141	New York, USA	2006	C	Invasive	11	11	20519285
α14	Bavaria, Germany	1999-2000	NG-*cnl*	Carriage	53	53	18305155
M17062	Minnesota, USA	2008	NG	Carriage	198	198	N/A
M15293	Georgia, USA	2006	NG (B)	Invasive	32	32	20519285
M16917	Illinois, USA	2007	NG	Invasi□e	11	11	N/A
α153	Bavaria, Germany	1999-2000	29	Carriage	60	60	18305155
α275	Bavaria, Germany	1999-2000	W135	Carriage	22	22	18305155
M13519	New York, USA	2005	W135 (C)	Invasive	11	11	N/A
M17661	Michigan, USA	2008	W135 (C)	Inva□ive	11	11	N/A
M18774	Florida, USA	2009	W135	Invasive	11	11	N/A
NM9261	□urkina Faso	2002	W135	Invasive	11	11	20519285
M11791	New York, USA	2003	Y	Invasive	23	23	N/A
M14900	Oregon, USA	2006	Y	Invasive	1625	23	N/A
M20899	California, USA	2009	Y	Invasive	1624	167	N/A

[a] Origins are based on literature searches for each genome or based on unpublished notes at Centers for Disease Control and Prevention.

[b] Serogroup. Sgs are defined as a result of the SASG test, and if PCR had a differing result its resulting serogroup is in parentheses (Mothershed *et al.*, 2004). *cnl*: the capsule locus is nonexistent and no capsule is expressed.

[c] Sequence Type (ST), Clonal Complex (CC).

[d] PubMed ID of the genome announcement. N/A: not applicable; not published yet.

Before these genomes had been completed, it was well known that many genes were shared using DNA uptake sequences (DUSs) as markers for neisserial origin (Goodman & Scocca, 1988; Smith *et al.*, 1999). Nearly 2000 of these occur in each meningococcal genome, and they function to incorporate flanking sequences through homologous recombination (Linz *et al.*, 2000). Another avenue of inducing homologous recombination is through insertion sequence (IS) elements (Mahillon & Chandler, 1998). IS elements in the meningococcal genome are on the order of 1kb and code for transposase activity. Pairs of IS elements can therefore act as a cassette and transpose flanking genomic regions to cause whole genomic rearrangements (Kawai *et al.*, 2006). Furthermore, simple repeat sequences such as dRS3 have been correlated with inversions (Schoen *et al.*, 2008).

PROTEIN

Many proteins are important foci for epidemiological, mechanistic, and immunological studies. Epidemiological studies are as described in the previous Population Structure section.

Mechanistic studies give insight into how the meningococcus colonizes and invades a human host. The line between colonization and invasion is blurred. Many adhesins such as pili are known to aid in epithelial adhesion between *N. meningitidis* and the human epithelium. In other words, these proteins are meant to help in colonization and not invasion. Seemingly by accident in symptomatic patients, these proteins begin helping in invasion. It is well-known that pili are important for invasion (McGee *et al.*, 1983). So therefore, a major question in invasion mechanism studies is, when is a protein considered invasive or just a way to inhabit a human host peacefully (Meyers *et al.*, 2003; Schoen *et al.*, 2008)? This question has not been fully answered and requires a better understanding of the mechanisms of invasion and pathogenicity.

Also studies elucidate how the meningococcus interacts with the immune system. One benefit of understanding this interaction would presumably be to make an all-inclusive meningococcal vaccine—only four out of the major five virulent serogroups are targeted by vaccines today, and ideally targeting all possibly-virulent serogroups is the goal. Therefore targeting a protein or set of proteins would be the best way to make a universal vaccine. Many proteins especially Opa and PorB (Estabrook *et al.*, 2004), PorA (Jarva *et al.*, 2005), factor H-binding protein (fHBP) (Fletcher *et al.*, 2004; Madico *et al.*, 2006; Masignani *et al.*, 2003), NadA (Litt *et al.*, 2004), and NHBA (previously GNA2132) (Lucidarme *et al.*, 2009) have been known to elicit an immune response.

Creating a universal vaccine was jump-started by a method called reverse vaccinology where a genome – the entire set of genes an organism can make – can be analyzed for vaccine potential (Rappuoli, 2001). The best qualities of a vaccine protein target are to be expressed outside the cell, to be conserved and not variable, and to be visible to the immune system (Pizza *et al.*, 2000).

Currently there are two major companies proceeding with the creation of a vaccine. Novartis is creating a vaccine based on five antigens, based on their reverse-vaccinology studies (Jacobsson *et al.*, 2009). These genome-derived neisserial antigens (GNAs) are a fusion of fHBP variant 1 and GNA2091, a fusion of NHBA and GNA1030, and NadA. The protein fHBP is a cofactor of factor I and aids the bacterium in avoiding complement-based attack by the immune system (Schneider *et al.*, 2006). NHBA is not well characterized but has some relevance to serum resistance, binds to heparin, and is a possible adhesin (Lucidarme *et al.*, 2009; Serruto *et al.*, 2010). It seems that GNA2091 plays a role in stress responses (Seib *et al.*, 2010). GNA1030 has not been characterized much yet but it seems as though efforts into understanding it are underway (Seib *et al.*, 2010). NadA is an invasin and adhesin which promotes adhesion to

epithelial cells (Capecchi *et al.*, 2005). A sixth ingredient is the outer membrane component PorA P1.7-2,4, which was a specific protein found in the MeNZB campaign against a 14-year epidemic in New Zealand (Sexton *et al.*, 2004a; Sexton *et al.*, 2004b). By having a diverse accompaniment of proteins, this vaccine will not allow for many escape mutants and will be effective in preventing meningococcal disease.

Another company Wyeth is approaching the problem by using one protein fHBP (Fletcher *et al.*, 2004; Masignani *et al.*, 2003). The population structure of fHBP has been analyzed without regard to MLST or other typing methods (Murphy *et al.*, 2009). In this way, most of the different variants of fHBP could be incorporated into the vaccine and therefore would be protective against most isolates. All isolates contain fHBP, and if the vaccine contains most variants of the protein, then it will prevent meningococcal disease.

There have been other vaccine attempts (Sadarangani & Pollard, 2010), many of them successful, but as of today only the Novartis and Wyeth vaccines are coming close to being universal. Presently, the most effective vaccines are those against the capsules of isolates belonging to serogroups A, C, Y, and W135 (Popovic & Ajello, 2003).

CAPSULE

The meningococcus has a group II polysaccharide capsule, which helps protect it from the immune system and from the environment (Costerton *et al.*, 1981; Swain & Martin, 2007). Meningococcal capsules are composed of polysaccharides and contain 95% water (think of a very watery gelatin mix). Therefore if it is thick enough, it is difficult for a phagocyte to ingest it or to destroy it with a lysosome. In addition it is protective against the environment. Meningococci do not live anywhere except in the human nasopharynx, but there is some assumed time that it must spend outside of the host before reaching the next host. The capsule prevents desiccation in this instance.

There are 13 serogroups: A, B, C, D, 29E, H, I, K, L, W135, X, Y, and Z (Ashton *et al.*, 1983). However, serogroup D might exist merely as an artifact of serogrouping assays (Popovic & Ajello, 2003; personal communication, Jennifer Dolan Thomas). Each serogroup differs in its molecular composition (Table 1.4). The differences between each serogroup include its choice of sugar monomer derivative; modifications such as acetyl groups; and linkages between sugars (Frosch & Vogel, 2006).

Table 1.4. Molecular composition of serogroups.

Serogroup D has been removed, as there is no chemical or genetic data available and also because it likely exists merely as an artifact of nonspecific antisera to serogroup C. Adapted from (Dolan-Livengood et al., 2003; Frosch & Vogel, 2006).

Serogroup	Structure	Genes
A	→6)-α-D-Man*p*NAc-(1→OPO_3→	*sacA-sacD*
B	→8)-α-D-Neu*p*5Ac-(2→	*synA-synD*
C	→9)-α-D-Neu*p*5Ac-(2→	*synA-synC*, *synE*
29E	→3)-α-D-Gal*p*NAc-(1→7)- β-D-KDO*p*-(2→	*cap29eA-cap29eH*
H	→4)-α-D-Gal*p*-(1→2)-Gro-(-3→OPO_3→	
I	→4)-α-L-Gul*p*NAcA-(1→3)- β-D-Man*p*NAcA(→	
K	→3)- β-D-Man*p*NAcA-(1→4)- β-D-Man*p*NAcA-(1→	
L	→3)-β -D-Glc*p*NAc-(1→3)- β-D-Glc*p*NAc-(1→3)-α-D-Glc*p*NAc-(1→OPO_3)→	*lbcA-lbcC*
W135	→4)-α-D-Neu*p*5Ac-α-(2→6)-α-D-Gal-(1 →	*synA-synC*, *synG*
X	→4)-α-D-Glc*p*NAc-(1→OPO_3 →	*xcbA-xcbC*
Y	→4)-α -D- Neu*p*5Ac-α-(2→6)-α-D-Glc-(1→	*synA-synC*, *synF*
Z	→3)-α-D-Gal*p*NAc-(1→1)-Gro-(3→OPO_3→	*capZA-D*

The capsule is genetically coded by the *cps* complex and can be divided into regions D', E, C, A, D, and B (Frosch *et al.*, 1989; Tzeng *et al.*, 2005). More important though are the operons at regions C and A which respectively code for transport and synthesis of the capsule. They are separated by a short intergenic region 134 nt long (Swartley *et al.*, 1997). The operon at region C codes for genes *ctrA-ctrD* which code for an ABC transporter. This transporter is also aided by the genes *ctrE* and *ctrF* in region B (Tzeng *et al.*, 2005) and also *ctrG* (Hobb *et al.*, 2010). Synthesis of the capsule is performed by region A which is an operon that contains *synA-synD* (Swartley *et al.*, 1996). The first genes *synA-synC* are responsible for creating enzymes that manufacture the monomer sugar. The last gene is a polymerase and is responsible for linkages. The *synA-synC* genes are homologous between serogroups that share the same sugar (*e.g.* serogroup B and C have the same modified sialic acid), but *synD* is specific to each serogroup. Because of the diversity of *synD* genes and the differences in sugars, the naming convention varies. Names for each of the *synA-synC* genes can be found in Table 1.4. The *synD* naming

convention is by serogroup and so in serogroup A the genes are named *sacA-sacD* (Syn: *mynA-mynD*), and in serogroup C the genes are named *synA-synC* and *synE* (syn: *siaA-siaD*$_C$). Because different serogroups have wholly different sugars, the *synD* genes are not considered alleles, but different genes. For consistency throughout this thesis, save chapter 8, the synthesis genes will be named *synA-synD.*

As a result of the plasticity of the meningococcal genome, a phenomenon called capsule switching can occur (Swartley *et al.*, 1997). Whole sections of the genome, on average 1.1 kb (Jolley *et al.*, 2005), can be homologously recombined. Consequently, if a section of the capsule synthesis region is recombined then the serogroup can switch. This phenomenon was first discovered in 1997 (Swartley *et al.*, 1997) and confirmed in 2000 (Vogel *et al.*, 2000). Because serogroup and ST can be correlated, it can now be said that when an isolate belongs to a serogroup not correlated with an ST, there is evidence of capsule switching (Harrison *et al.*, 2010). It occurs more often than expected; 4.3% of all isolates were determined to have arisen as a result of capsule switching in a 2010 study (Harrison *et al.*, 2010).

Another phenomenon related to the capsule is nongroupability. An isolate is considered nongroupable (NG) if it has no capsule, if it belongs to more than one serogroup, or if the isolate is untestable, *i.e.* it cannot be placed in any one group (Mothershed *et al.*, 2004). The gold standard to test for serogroup is through the slide agglutination serogrouping (SASG) test. Isolates are placed into separate whole-cell ELISA tests, one for each serogroup. If no reaction takes place then it belongs to no serogroup, and likely, placing it into human serum will kill the cells due to the immune system (Dolan-Livengood *et al.*, 2003). If the isolate agglutinates in the presence of saline and not antibodies then it is an autoagglutinator and is nongroupable. If the SASG test results in more than one positive, then the isolate is nongroupable because it belongs to more than one group. Another serogrouping test is a PCR test. Sites on the *synD* genes have

been targeted for standard primers for qualitative PCR. Each serogroup's set of primers is used on an isolate, and if there is a positive result in the PCR, it is positive for that serogroup. The PCR test adds quality to serogrouping but also adds another phenomenon where the SASG test and PCR test disagree (Mothershed *et al.*, 2004). If the results disagree, it is called a discrepant result (DR).

Several vaccines target polysaccharide capsules. The two most notable polysaccharide vaccines are tetravalent for serogroups A, C, W135, and Y. The first, MPSV4, is composed only of the polysaccharide, purified from meningococcal lysate. It is still widely used for children 2-10 years of age and persons over 55 (Bilukha & Rosenstein, 2005). The other notable tetravalent vaccine, MCV4, is licensed for people 20-55 years old and is recommended for those 11-55 (Bilukha & Rosenstein, 2005). MCV4 is a conjugate vaccine composed of the 4 types of polysaccharide and diphtheria toxoid. Conjugate vaccines are theoretically better because they induce a T-cell response which creates memory T-cells; however, preliminary data show that MPSV4 and MCV4 both provide the same level of immunogenicity (personal communication, Jessica MacNeil).

DESCRIPTION OF THIS THESIS

This thesis is an answer to many of the questions and problems above. First, MLST and other DNA-based typing methods are cumbersome. I have devised a platform which will be described in **chapter 2** that makes DNA-based typing much easier for the experimentalist. These tools, dubbed The Meningococcus Genome Informatics Platflorm (MGIP), are currently being used at the Centers for Disease Control and Prevention (CDC) and other world-class laboratories for meningococcal surveillance and are replacing the software Sequence Typing Analysis and Retrieval System (STARS) (Chan & Ventress, 2001). In the last year, the CDC Influenza Division Sequencing Activity Laboratory has begun a collaboration with me to modify MGIP to

accompany a new sequence typing method called Small Target Reassortant Screen (STaRS). STaRS is used by the Influenza Division to aid in worldwide surveillance of influenza much like MLST is used to observe *N. meningitidis*. To analyze STaRS data, I created the Influenza Genomics Platform (InGen). I will discuss InGen further in **chapter 3**.

Aside from epidemiology, a major question persists in the study of meningococci (Meyers *et al.*, 2003). Why are most strains carried asymptomatically while only a small percentage of the population causes acute disease? *N. meningitidis* is most frequently found to be asymptomatically carried (Maiden & Caugant, 2006). Carriage studies reveal that about 10% of healthy individuals are carriers of *N. meningitidis* (Claus *et al.*, 2005; Dolan-Livengood *et al.*, 2003). Thus, *N. meningitdis* strains can be phenotyically divided according to their pathogenic potential: invasive versus carriage strains. Invasive strains may be asymptomatically carried for some time, but they also have the potential to cause invasive disease. The vast majority of meningococcal meningitis outbreaks are caused by a limited number of invasive strains. Carriage strains, on the other hand, are not invasive and thus do not have a large potential to cause disease. Understanding the genomic differences between invasive versus carriage strains may provide fundamental insight into the mechanisms of virulence in *N. meningitdis*. Although this question has not been answered experimentally yet, I have narrowed the search considerably as shown in **chapter 6**. Owing to the fact that the presence or absence of any gene in the meningococcal genome cannot unambiguously determine virulence (Schoen *et al.*, 2008), I performed a genome-wide analysis to detect single nucleotide polymorphisms (SNPs). However in order to perform whole-genome analyses, a genome assembly annotation pipeline had to be developed. At the time of this thesis proposal in October 2008, only four invasive meningococcal genomes had been sequenced (Bentley *et al.*, 2007; Parkhill *et al.*, 2000; Peng *et al.*, 2008; Tettelin *et al.*, 2000), and this number had to grow before performing representative

studies. In **chapter 4**, I describe how I led a team of bioinformaticians to create a genome assembly and annotation pipeline called the computational genomics pipeline (CG-Pipeline). CG-Pipeline produces annotated genomes from 454 pyrosequencing runs (Margulies *et al.*, 2005). Owing to the next-generation technologies and to CG-Pipeline, the time it takes to produce an annotated bacterial genome from the nasopharynx has been reduced from years to days.

Although the final GenBank file outputted from CG-Pipeline can be viewed by a biologist, it was a natural next step to produce a genome browser. This genome browser would be capable of interactively navigating through a genome, with added features such as searching. In **chapter 5**, I describe how I led a team of bioinformaticians to produce *Neisseria* Base (NBase) to achieve such a purpose.

Using these finalized genomes I was able to address another open question: how are some isolates nongroupable, and what do they look like on a genetic level? In **chapter 7**, I show how the *cps* complex appears on a genetic level in such a genome.

Therefore this thesis describes how I created a tool to aid in epidemiological surveillance of *N. meningitidis* and influenza; how I created a genome annotation pipeline as the counterpart to next-generation sequencing technologies; how I compared several genomes to find markers for invasiveness in *N. meningitidis* which will be targets of future experimental investigation; and how I characterized a nongroupable genome. These studies are currently aiding in public health efforts to fight *Neisseria meningitidis* and will possibly aid in fighting other infectious agents worldwide.

CHAPTER 2

MENINGOCOCCUS GENOME INFORMATICS PLATFORM: A SYSTEM FOR ANALYZING MULTILOCUS SEQUENCE TYPING DATA

ABSTRACT

The Meningococcus Genome Informatics Platform (MGIP) is a suite of computational tools for the analysis of multilocus sequence typing (MLST) data, at http://mgip.biology.gatech.edu. MLST is used to generate allelic profiles to characterize strains of *Neisseria meningitidis*, a major cause of bacterial meningitis worldwide. *N. meningitidis* strains are characterized with MLST as specific sequence types (ST) and clonal complexes (CC) based on the DNA sequences at defined loci. These data are vital to molecular epidemiology studies of *N. meningitidis*, including outbreak investigations and population biology. MGIP analyzes DNA sequence trace files, returns individual allele calls and characterizes the STs and CCs. MGIP represents a substantial advance over existing software in several respects: 1) **ease of use** - MGIP is user friendly, intuitive and thoroughly documented; 2) **flexibility** - because MGIP is a website, it is compatible with any computer with an internet connection, can be used from any geographic location, and there is no installation; 3) **speed** - MGIP takes just over one minute to process a set of 96 trace files; and 4) **expandability** - MGIP has the potential to expand to more loci than those used in MLST and even to other bacterial species.

INTRODUCTION

Epidemiological surveillance of *N. meningitidis* necessitates molecular typing. Standard methods for molecular typing include, but are not limited to, restriction fragment length polymorphism (Arreaza & Vázquez, 2001), pulsed field gel electrophoresis (Achtman & Morelli,

2001), multilocus sequence typing (MLST) (Jolley, 2001; Jolley *et al.*, 2006), and *porA*, *porB*, *and fetA* typing (Sacchi *et al.*, 1998; Sacchi *et al.*, 2000; Thompson *et al.*, 2003). MLST is the most modern and widely used of these approaches, and it provides an unambiguous method for typing bacterial strains (Maiden *et al.*, 1998). In MLST, specific regions of seven housekeeping genes are sequenced, their alleles are determined, and then the allele calls are concatenated to produce a profile called the sequence type (ST), which may then be grouped into a larger population called a clonal complex (CC). MLST analysis used in conjunction with the molecular typing methods listed above can provide evidence of possible genetic and epidemiological relatedness of strains identified during outbreak investigations and routine surveillance (Enright & Spratt, 1999).

Epidemiological surveillance laboratories world-wide perform MLST using PCR and Sanger sequencing. Standard primers are used to amplify each of the seven loci, and the PCR fragments are then characterized using dye-terminator sequencing. The resulting trace files are interpreted by a computer or a human and are converted into unambiguous sequences (base calling). Some computer programs that will make base calls are Phred (Ewing & Green, 1998; Ewing *et al.*, 1998) and SeqMan (DNASTAR SeqMan Pro, Madison, WI). If there is more than one sequence read per locus, then those sequences must be assembled to generate a single consensus sequence. Computer programs that can perform assembly include Phrap (Gordon *et al.*, 2001) and SeqMan. The last step in MLST analysis is to determine the allele of the gene by comparing the consensus sequence of the trace files to a database of known allele sequences. In MLST, even a single nucleotide difference is sufficient to define a new allele and thus the comparison between the consensus sequence and the allelic database must be unambiguous.

The current standard software for MLST analysis, STARS (http://sara.molbiol.ox.ac.uk/userweb/mchan/stars/), is no longer supported by the original

programmers, only runs on UNIX/Linux systems, and has performance and usability issues. STARS is included in distributions of Bio-Linux, but comprehensive and detailed instructions for its set up are necessary (http://pubmlst.org/software/bio-linux/stars/config/). A commonly used alternative to STARS is to make base calls, assemble sequences manually and then use the BLAST (Altschul *et al.*, 1997) interface at pubmlst.org (Jolley *et al.*, 2004) to ascertain the identity of the MLST alleles. MLST users and laboratories may use a variety of packages to analyze trace files and make base calls, including SeqMan, MEGA (http://www.megasoftware.net), BioNumerics (Platt *et al.*, 2006) (http://www.applied-maths.com/bionumerics/bionumerics.htm) and the CLC MLST module (http://www.clcbio.com/index.php?id=1018). These alternatives, while viable, either require programming expertise to provide expanded capabilities, accessibility to a Linux system, or are prohibitively expensive. In addition, most of these alternatives represent piecemeal and tedious approaches that require a substantial dedication of time and resources. For all of these reasons, MLST analysis can be burdensome for laboratories in developing countries and/or in laboratories with fewer resources for personnel and computational support. To address these problems, we have developed an integrated suite of MLST analysis tools available at the Meningococcus Genome Informatics Platform (MGIP).

MGIP presents several key advantages over currently existing analytical methods: 1) **Ease of use -** MGIP runs as a web server, is designed to be user friendly, intuitive, and is thoroughly documented. The documentation is on the website itself and covers any topic the user might need to see. If in any case the documentation is not sufficient, there is a conspicuous link to contact the lab for help; 2) **Flexibility -** MGIP is compatible with any client computer or operating system and has been tested using Microsoft Internet Explorer, Mozilla Firefox, Safari, and Google Chrome. Much of this flexibility is given by the cross-browser compatible JavaScript frameworks Prototype and Scriptaculous; 3) **Speed -** MGIP has been shown to take about 1

minute per set of 96 sequence trace files, more than 5 times faster than STARS. The speed of MGIP can be attributed to the fast constituent programs Phred, Phrap, and BLAST as well as the server which is an eight-core machine. In addition, MGIP's ability to process multiple loci concurrently is a considerable advantage in comparison to other MLST analysis tools, which can only process one locus at a time; 4) **Expandability –** Currently MGIP can analyze sequences from over 15 loci, which substantially increases the resolution of sequence typing. MGIP has the potential to include unlimited loci and has the capacity to include other organisms. We refer to MLST analysis with additional alleles as MLST+. Additional loci can be added in one of two ways: either by the administrator of MGIP or by an individual user. If a locus is added by the administrator, then all users can use the new locus database as both a BLAST database and for analyzing new trace files. If a user adds a database, it will be only visible to that user.

MLST+ ANALYSIS WORKFLOW

The workflow and programs underlying MLST+ analysis with MGIP are shown in Figure 2.1. Users of MGIP first sequence a set of loci to be used in MLST+. The loci that MGIP can process by default are shown in Table A.1. Current protocols for *N. meningitidis* surveillance laboratories usually yield sequence data from 96-well plates, but the number of wells or traces does not affect the MGIP workflow.

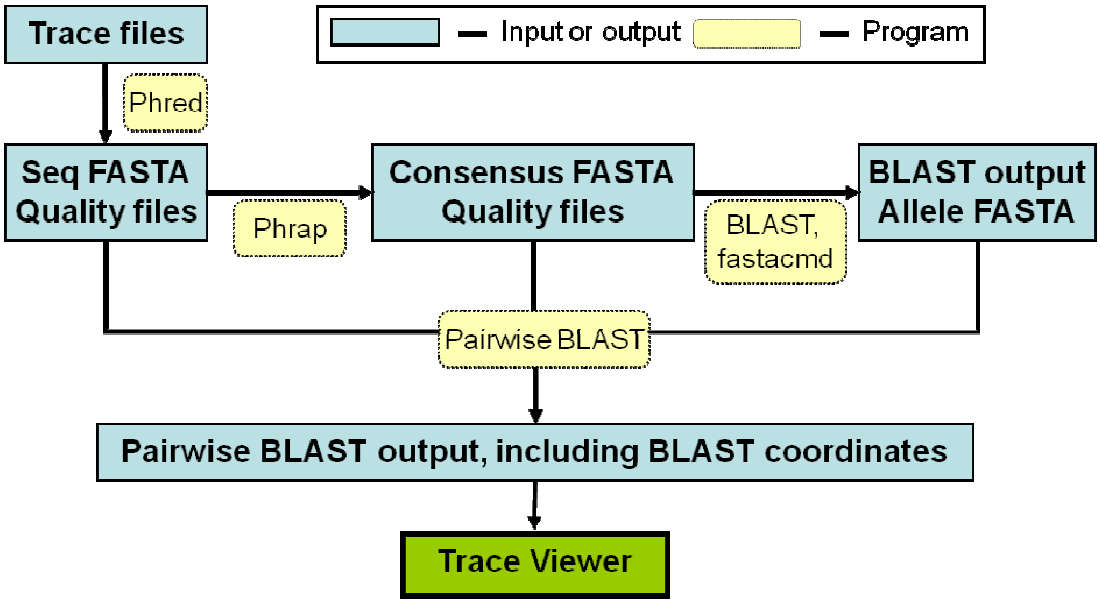


Figure 2.1. MLST+ workflow on MGIP. Users upload trace files to the MGIP server for analysis. First, Phred makes base calls on each trace to produce a sequence FASTA file and a quality file. Next, Phrap aligns and produces a consensus sequence FASTA file and other associated files. BLAST is then used to match the consensus sequence against a database of known MLST+ alleles. Allelic FASTA files are extracted from the database using fastacmd and individually aligned to the consensus sequences to determine coordinates, mismatches, and indels using pairwise BLAST. Alignments between consensus sequences, called allelic sequences, and underlying trace files are displayed using the trace file viewer. The trace file viewer can be used to manually edit consensus sequences based on the aligned trace files (See Figure 2.3).

MGIP takes two files as input. The first file is a zip file of every trace received from the sequencing machine in one session. The second file is a mapping spreadsheet that assigns the following properties to each trace file: strain name, sequence typing method, locus, and primer.

After these two files are submitted to MGIP, the trace files undergo processing: 1) Phred makes base calls on each trace file; 2) Phrap assembles groups of trace files to make a consensus sequence for each strain/locus; and 3) MGIP uses BLAST against a database of known MLST+ alleles to determine the allelic identity of the consensus sequence. BLAST results that do not have perfect matches in the database are flagged (perfect matches have 100% identity, 100% subject coverage, and no indels). These flags are shown when viewing results and call attention to possibly novel or inaccurate results. MGIP also includes a trace viewer that shows alignments of consensus sequences, allele calls and underlying trace files. The trace viewer can be used to manually edit consensus sequences based on the aligned traces. The results of MGIP analyses are public unless users are registered and logged in at the time analysis is performed.

MGIP EASE-OF-USE AND UNIVERSAL COMPATIBILITY

One of the goals for the development of MGIP was to make a system that is simple and convenient to use. This goal is achieved via 1) an intuitive user interface, 2) operating system and browser cross-compatibility, and 3) thorough documentation. These features are particularly relevant in the developing world where technical help and systems support may be scarce, but they are also applicable to scientists everywhere who do not wish to devote substantial resources, in both time and hardware, to computation.

As opposed to STARS, which requires machines running Linux, MGIP can be used with any operating system since it is run on a server with a web browser based client interface. MGIP is compatible with most standards-compliant web browsers because it largely conforms to the worldwide standards given by the World Wide Web Consortium (W3C). The Prototype and

Scriptaculous frameworks, which are thoroughly tested on many browsers for compatibility, were used in the development of MGIP to ensure JavaScript compatibility. MGIP has been tested on Microsoft Internet Explorer, Mozilla Firefox, Safari, and Google Chrome, which together account for almost 99% of all web browsers in use (http://marketshare.hitslink.com/, last accessed November 1, 2008).

USING MGIP

USER ACCOUNTS

MGIP was developed to allow users to upload private or sensitive data. Therefore, MGIP has a user management system, with few administrators and many users. Each user inherently has all of his or her data and user information privatized so that no other user can access them. To this end, MGIP employs standard web server security measures including MD5 password encryption and the use of a firewall (http://www.faqs.org/rfcs/rfc1321). For data that is not sensitive, or for scientists who do not wish to use individual accounts, there is a default public user setting with full functionality except data privatization.

UPLOADING TRACES AND RUNNING ANALYSES

Users upload trace files to MGIP for typing analysis. The user must create a zip file from all sequence trace files and a mapping spreadsheet file which identifies each trace's strain, sequence typing method, locus, and primer. This spreadsheet is crucial for assembling the correct sequences together and provides names to each of the final results. An automatic spreadsheet generator is available which will generate the spreadsheet.

VIEWING RESULTS BY SET

Fully automated analysis by MLST+ produces results for sets of traces which are viewed on the main page (Figure 2.2). For each set, the allele calls for each locus are displayed. For each locus, there is a submenu with options to 1) view BLAST results; 2) view the consensus

sequence and quality scores as given by Phrap; 3) download all files involved in the MLST+ analysis workflow; and 4) view trace files and edit the consensus sequence, thus allowing the user to manually adjust the results. BLAST results are reported in default format and for each hit show the allele names, bit scores and e-values along with pairwise query-hit sequence alignments. Consensus sequences are shown in FASTA format, with each nucleotide shown in a color corresponding to a range of quality scores. Loci that yield ambiguous results from the MLST+ analysis workflow have flags on the results screen that show the user where to intervene.

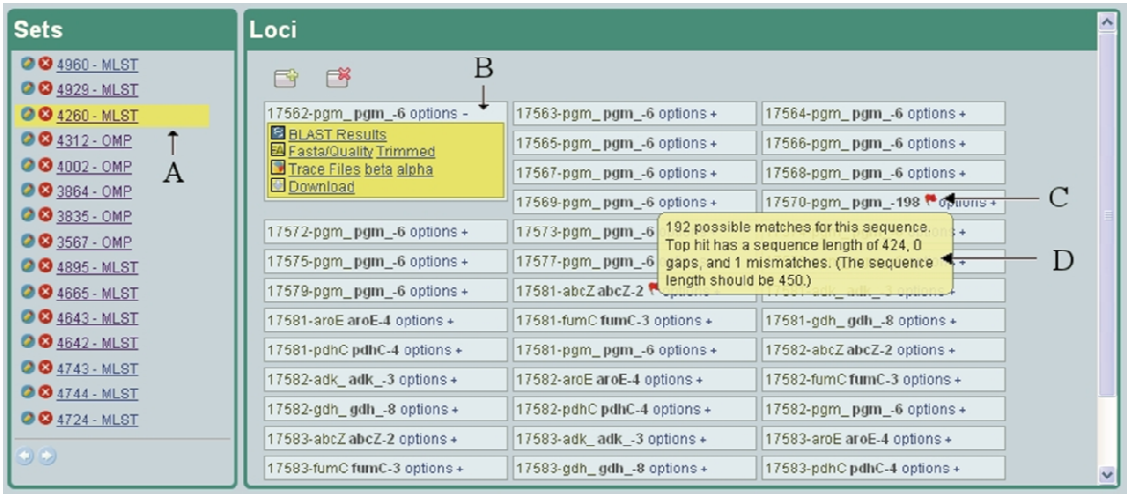


Figure 2.2. Viewing MGIP+ results. (A) The user can select a set of results to view. For each set, all strain/loci and their allele calls are shown. (B) Options are shown for each strain/locus that allow the user to view more details. (C) When an allele call is not a perfect match, a flag appears. (D) On mouseover (when the mouse pointer hovers over the flag), a message giving information as to why it is not a perfect match appears.

To view trace files, we have developed a trace viewer Java applet that displays the alignment of all traces involved in the assembly process, the consensus sequence, and the allelic sequence (Figure 2.3). The applet automatically marks all discrepancies between the trace file base calls, the consensus sequence, and the most similar allele, to facilitate scanning for inconsistencies. Trace amplitudes can be adjusted individually, and although the traces are trimmed to show only the aligned regions, there is an option to view trimmed edges. The consensus sequence can be edited by changing, adding, or deleting bases using the trace viewer. Once manual editing is completed, the workflow analysis starting with BLAST can be iterated so that the main page results are updated.

Figure 2.3. The trace viewer and editor applet. The consensus sequence acts as a backbone when aligning the allelic sequence and the traces. The applet tools allow users to 1) alter the amplitude of the traces, 2) edit the consensus sequence, 3) insert/delete consensus sequence nucleotides, 4) undo/redo any action, and 5) save a modified consensus sequence. Sequences of interest are embedded below the applet so that they can be copied and pasted.

VIEWING RESULTS IN THE STRAIN TABLE

The results of MLST analyses are also displayed in the strain table, which shows the allelic profile, ST and CC for each strain that has been analyzed (Figure A.1). For each individual

user account, the strain table is automatically and continually populated with the combined results of all sets that have been analyzed. If fewer than seven known alleles have been unambiguously characterized for any given strain, a list of all possible STs and CCs is shown.

REFERENCE PAGES

To aid in data analysis, all reference data in the MGIP database has been made transparent and available on the reference pages. The ST reference page allows the user to type in an ST number and to retrieve the alleles associated with that ST. Alternatively, a CC number can be input to retrieve all STs associated with it. The locus reference page shows every locus that can be analyzed using MGIP. For each locus, the sequence typing method and the length of the allele are shown; all sequences in the locus databases can be downloaded. The locus reference page also has a BLAST interface, which accepts one or multiple FASTA query entries for comparison against the locus database.

PERFORMANCE VALIDATION

MGIP was compared against two other methodologies to validate performance in terms of both sensitivity and speed (Table 2.1). Sensitivity is defined as:

$$Sn = \frac{TP}{TP + FN} \quad \textbf{1}$$

where *Sn* is sensitivity, *TP* is the number of true positives, and *FN* is the number false negatives. In this study, a true positive is defined as an unambiguously identical match from the trace file(s) to the allelic database, and a false negative is defined as no match when there should be one. Specificity cannot be measured because all methodologies in this study filter out false positives before they are reported. Speed is calculated simply as the time elapsed from upload to the end of sample processing.

Table 2.1. MGIP is more sensitive and faster than other commonly used methods.

MGIP vs STARS[a]				
	TP[c]	*FN*[d]	*Sn*[e]	speed (seconds)[f]
MGIP	660	18	97.4%	63±0.58
STARS	653	35	94.9%	323±30

MGIP vs SeqMan method[b]				
	TP[c]	*FN*[d]	*Sn*[e]	speed (seconds)
MGIP	323	6	98.2%	75±2
SeqMan	319	8	97.6%	1520±173

[a] For the MGIP vs. STARS comparison, 17 sets of MLST data were tested which were composed of trace files over 691 strain/loci. The speed test was performed on 3 randomly selected sets, composed of 126 strain/loci.

[b] For the MGIP vs. SeqMan method comparison, 10 sets of *fetA* were tested in the SeqMan comparison, totaling 331 loci. The speed test was performed on 3 randomly selected sets, composed of 103 strain/loci.

[c] *TP* is true positives.

[d] *FN* is false negatives.

[e] *Sn* is sensitivity.

[f] Speed is shown as an average per trace file set (84 traces in the STARS comparison, 96 traces in the SeqMan comparison), plus or minus standard deviation. The speed tests were performed over a 1 Gigabyte per second network connection and therefore the upload time was negligible. However, the upload time from a slower connection will understandably increase the time to process a set of trace files. Approximate times for uploading a set of traces is given in Table A.2.

The first methodology compared to MGIP was STARS, which is the current standard for MLST analysis. For comparison to STARS, 17 MLST sets were analyzed totaling 691 strain/locus combinations. MGIP showed 97.4% sensitivity compared to 94.9% sensitivity for STARS (Table 2.1). In addition to being more sensitive, MGIP is also substantially faster than STARS. On average, MGIP finished analyzing a set of 84 trace files in 63 seconds compared to 323 seconds for STARS (Table 2.1).

The second methodology compared to MGIP was the "SeqMan method," where a consensus sequence is created from trace files using SeqMan and used as a query in the Pubmlst BLAST interface. SeqMan is used when non-standard MLST alleles are being analyzed, and requires substantial manual intervention by the user. Ten sets totaling 331 *fetA* traces were analyzed for the comparison of MGIP to the SeqMan method. MGIP showed 98.2% sensitivity compared to 97.6% for the SeqMan method (Table 2.1). MGIP showed an even greater relative increase in speed over SeqMan. MGIP completed the ten *fetA* trace sets in 75 seconds compared to 1520 seconds for the SeqMan method (Table 2.1).

FURTHER DEVELOPMENT OF MGIP

There are several lines of further development of MGIP planned. MGIP allows for the discovery of novel alleles and/or STs, which cannot be named or curated until they are sent to one of the central repositories of MLST data such as Pubmlst. We have been collaborating with the developers of Pubmlst to design an application programming interface that will allow MGIP users to directly submit new alleles and new STs. In addition to *N. meningitidis*, there are many more bacterial pathogens that are analyzed using MLST and MGIP will add the capacity to analyze additional organisms in the near future. The MGIP website is being translated to other languages, starting with French, to facilitate collaboration with non-English speakers.

CONCLUSION

The web based design and implementation of MGIP helps to ensure that it stands alone among MLST analysis methods in terms of cost, speed of processing, ease-of-use, cross-compatibility and expandability. These features are particularly relevant to laboratories in the developing world, many of which may lack access to the level of computational infrastructure and support currently needed for MLST analysis. The use of simple web-based analytical platform should allow any investigator with Internet access to rapidly analyze his or her MLST data. Furthermore, MGIP is designed to be scalable to accommodate MLST+ analysis of multiple non-standard alleles. This feature should enable the expansion of current MLST based surveillance approaches.

The development of MGIP has been done in close contact with typing centers around the world to ensure that it will emerge as the global standard for MLST+ analysis. Labs that have been testing MGIP include the Meningitis laboratory of CDC in the USA, the Health Protection Agency in England, Martin Maiden's research group at the University of Oxford, and the National Institute for Communicable Diseases in South Africa. MGIP has been tested on over 1000 different strain/locus combinations, and the results show that it is 1-3% more sensitive and an order of magnitude faster than existing methods for MLST+ analysis.

ACKNOWLEDGEMENTS

The authors would like to thank Ahsan Huda and Maryanne Ku for their help in making the site more intuitive and easier to use. Also, many thanks go to Keith Jolley for his technical expertise and encouragement. Lastly, we would like to thank Troy Hilley for maintaining the MGIP server.

CHAPTER 3

THE INFLUENZA GENOMICS PLATFORM: A SYSTEM FOR ANALYZING SMALL TARGET REASSORTANT SCREEN DATA

ABSTRACT

The Influenza Genomics Platform (InGen) is a suite of computational tools for the analysis of Small Target Reassortant Screen (STaRS), at http://flu.biology.gatech.edu. STaRS is used to generate clade profiles to characterize strains of influenza, which has the potential to cause massive outbreaks. Furthermore InGen can utilize STaRS to display reassortment quickly and easily to the laboratorian.

INTRODUCTION

INFLUENZA

Influenza is a virus that belongs to the Orthomyxoviridae family. The genome is divided into eight linear segments of negative-sense single-stranded RNA and codes for ten genes. The epidemiologically relevant proteins that these ten genes code for are hemagglutinin (HA) and neuraminidase (NA). HA is the protein that recognizes and targets host cells, and it has a fusion domain to give access to the host cell. NA is the protein that allows the virus to be released from the host cell.

Influenza causes the flu whose symptoms include chills, fever, sore throat, muscle pains, severe headache, coughing, and fatigue. It is an airborne virus, usually spread by coughing or sneezing. Influenza spreads in seasonal epidemics and results in hundreds of thousands of deaths every year and up to millions in pandemic years. There have been many infamous

pandemics including the 1918 Spanish Flu, the H5N1 Avian Flu which is still ongoing, and the 2009 H1N1 pandemic.

To survey the spread of influenza worldwide, scientists have created typing methods. Influenza can be divided into genera, the major ones being A, B, and C. Influenza A is the most common source of human flu and can be further divided into serotypes (or subtypes). These types are based on the variant of HA and NA and some examples are H1N1, H1N2, and H5N1. However, even these subtypes are very diverse and so they are further subdivided into clades. For example, the strains related to the 2009 pandemic strain is one clade of the H1N1 serotype.

At times if a host is infected with two different strains, influenza can exchange segments of its 8-segment genome with other virions' segments. This process is called reassortment, and the new strain can be described as a reassortant strain. Therefore a person coinfected with H1N1 and H2N2 can produce the reassortant subtypes H1N2 and H2N1. This is one example of how influenza can evade the immune system.

CURRENT METHODS OF SURVEILLANCE

The CDC Influenza Division has been charged with observing and answering flu epidemics. Its current methods involve looking at the subtype and even the clade of the virus. The current system that CDC uses is called Ibis, which makes use of the Ibis T5000 universal biosensor platform (Sampath *et al.*, 2007). The Ibis methodology involves creating cDNA through reverse transcription and amplifying a small segment of the influenza genome. Next, the small target is sent though a mass spectrometer. Generically, this methodology is called RT-PCR/ESI-MS. The raw result of RT-PCR/ESI-MS is a mass spectral data plot. This plot can be analyzed and matched with one of several reference plots to determine the clade of the virus.

RT-PCR/ESI-MS is beneficial because it is somewhat quick and does not require much human intervention. Additionally it allows for quick detection of reassortant which would help advise

on which clades to vaccinate against in a flu season. However, RT-PCR/ESI-MS is on the whole, not beneficial because it is a method of changing digital nucleotide information into analog mass spectral data. Furthermore, it is difficult for laboratories around the world to standardize the results from this methodology. Remaining on the digital level and aligning nucleic acid information would be more beneficial because 1) it is unambiguous, 2) it is portable between laboratories, and 3) it is easily standardized between laboratories.

Therefore, the CDC Influenza Division's Sequencing Activity is moving away from Ibis and will be using a nucleic acid approach called Small Target Reassortant Screen (STaRS). Small targets in the genome covering all eight segments, as defined by PCR primer pairs, are chosen (Figure 3.1). Sanger sequencing is performed on these small targets which are typically less than 100 nt. The sequence is inferred from the raw data obtained from Sanger sequencing and is compared against a database of target variants. The closest variant, as defined by a low BLAST e-value, determines the clade that the query sequence belongs to. Additionally, if two sequences from the same genome match two different clades, then a reassortant event will be detected.

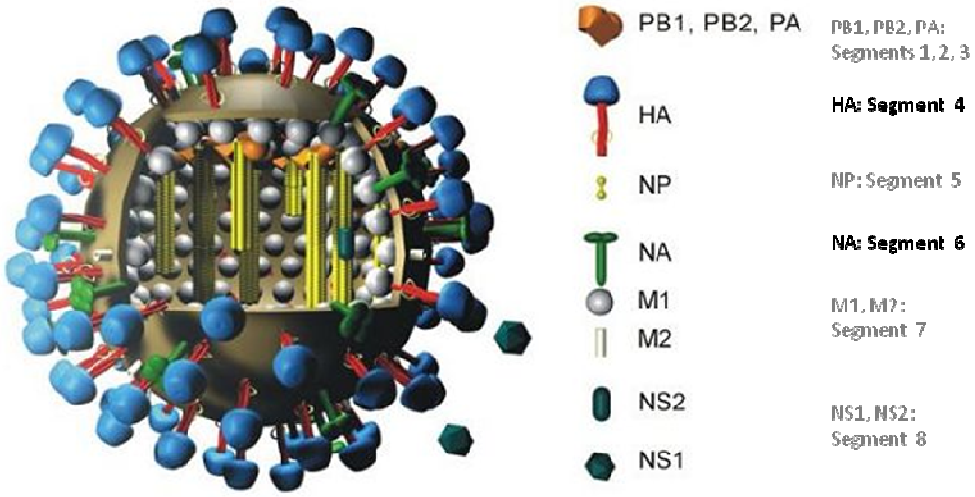


Figure 3.1. The influenza genome. The influenza genome is comprised of eight segments which are listed after the genes for which they encode. Segments 4 and 6 encode for the HA and NA genes which determine the subtype (e.g. H1N1). Figure adapted from Influenza Report, http://www.influenzareport.com/ir/images/virus.jpg

Even with these well-defined techniques, information still escapes CDC because it is not reported to CDC. There are many health laboratories around the U.S. and the world, each with their own records of influenza. However their results are not reported to CDC very often, which results in a national or global delay in influenza surveillance. As a result, it is difficult to predict which clades or even which subtypes might be circulating in a given time period. A consequence of this information lag is that vaccine preparation can go awry or can be delayed. Therefore some national or global data-sharing plan is necessary to combat influenza.

Fortunately our laboratory has created the Influenza Genomics Platform (InGen) which is the bioinformatic accompaniment to STaRS. InGen uses a digital approach to STaRS in that it 1) reads trace files, 2) assembles a consensus sequence at each target, and 3) performs BLAST to infer variant and profile calls. In this article, we will discuss how InGen has been developed to be a highly suitable auxiliary to worldwide usage of STaRS. InGen is available at http://flu.biology.gatech.edu.

INGEN FACILITATES STARS ANALYSIS

A COMPUTATIONAL APPROACH

STaRS analysis can be thought of as a workflow. A sequence from a genome must be read and combined with overlapping targets. The closest sequence from a reference database to each target identifies its clade/subtype/serotype. If sequences in a genome differ in origin, then the virus can be considered reassortant. For example, if targets from segments 1-7 come from one clade and segment 8 comes from another clade, then the virus is reassortant. Reassortment causes a heterogeneous genome, while other genomes are homogeneous.

InGen facilitates this workflow computationally (Figures 3.2 and 3.3). Raw Sanger sequence data that overlaps a target site for a particular strain is sent to InGen in a group. Each trace file is base-called to produce a digital sequence file and quality file. Next these sequences, in the

presence of their quality files, are assembled to produce a comprehensive target's sequence and quality file. This consensus sequence represents an overall region that is covered by the target sequences. Last using BLAST, these consensus sequences are compared against a curated reference database that represents each clade. There is one database per target. The target sequence is assigned a clade according to the top hit in the target database. The result is a clade call.

Each target's clade call is combined into a profile which is compared against a curated STaRS profile database. STaRS profiles are named using free text (*e.g.* H1N1pdm for the 2009 H1N1 pandemic strain). If the clade calls are from more than one STaRS profile then it is reassortant. However, if the clade calls are from the same profile, then no reassortment has taken place.

DATA SHARING

Much of STaRS data is sensitive, and so InGen has a user system (Table 3.1). Users may register and log in before analyzing data. Therefore, all user data is initially only available to the user who analyzed it. However, InGen also allows for groups. Sets of data, which are all data that are analyzed at one time and usually from a 96-well plate, can be added to a group.

A user can create a group by going to the manage groups menu. A link is there to create a group, and then a name and description must be filled in. After the group is created, the user becomes the group administrator. Users can be added to a group and can be given permissions: read, edit, add others, etc. Group administrators have *carte blanche* permissions. Also after the group is created, sets can be added. Sets that are associated with a group are visible to all members, as long as they have read permissions.

InGen has an application programming interface (API) which is a form of web services. An API is a method of accessing a web tool, or suite of tools such as those on InGen, and sending commands. For example, an API user would be able to send trace files and retrieve clade calls

from the command line. These API commands are thoroughly documented on the site using PHPDoc, which is similar to the standardized and well-known JavaDoc.

The disadvantage of using an API is that a user cannot experience the graphical interface that the website provides. However, the advantages are great to a user with programming experience. The core functionality of InGen is exposed when using the API, meaning that every function of the platform runs more quickly. The styling of the site and the JavaScript functionality is not loaded, which cuts down on load time significantly. Another major advantage is that a programmer could incorporate InGen functionality into a script.

An example of such a script would be one that finds clade calls for all trace files in a given directory and reports any reassortment. In this scenario, a laboratorian would run a sequencing machine and allow the results to be deposited into a folder upon completion. Early in the morning, this script would generate profiles for every strain and highlight those strains that are reassortant. The profiles would be written into a document, waiting for the laboratorian to start his/her day.

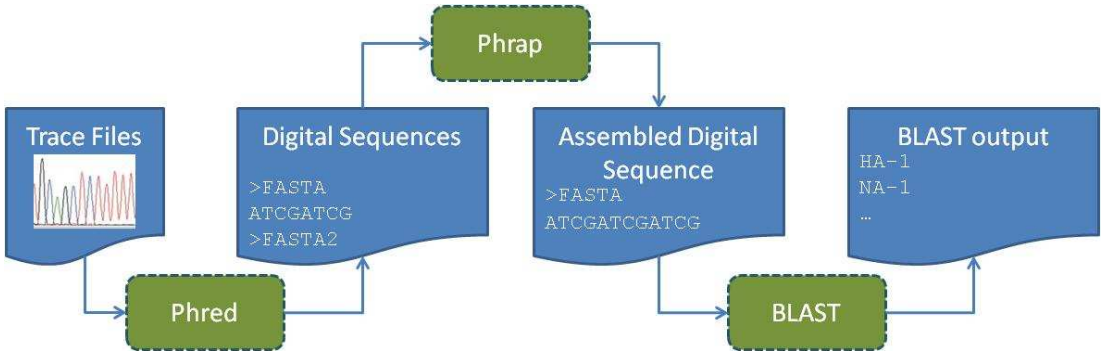


Figure 3.2. The InGen workflow. Users upload trace files to the InGen server for analysis. First, Phred (Ewing & Green, 1998; Ewing et al., 1998) makes base calls on each trace to produce a sequence FASTA file and a quality file. Next, Phrap (Gordon et al., 2001) aligns and produces a consensus sequence FASTA file and other associated files. BLAST (Altschul et al., 1997) is then used to match the consensus sequence against a target database with representative sequences for each clade. Allelic FASTA files are extracted from the database using fastacmd and individually aligned to the consensus sequences to determine coordinates, mismatches, and indels using pairwise BLAST. Alignments between consensus sequences, called target sequences, and underlying trace files are displayed using the trace file viewer. The trace file viewer can be used to manually edit consensus sequences based on the aligned trace files.

Figure 3.3. The InGen front page. From here, the user can access any InGen functionality.

DISCUSSION

InGen is a suite of tools that answers several issues in the influenza surveillance world. One problem is detection of reassortment. InGen can automatically detect all reassortment, given trace files of several targets in the influenza genome. Another problem is economical. InGen is free software that promises to be cutting edge and useful to any influenza typing laboratory. It is also reliable. Once data is sent to the InGen server, it is stored and analyzed independent of the laboratorian's desktop computer. These results can be downloaded at any time to view them.

InGen rests on a powerful computer. The server is a 2 quad core Intex Xeon 1.60 GHz with 8GB of RAM and four 500GB hard drives. The operating system is Ubuntu 8, and the server uses PHP 5.2, Mysql 5.0, and Apache 2. These specifications are crucial for its speedy analysis and data sharing duties.

Because InGen is cutting edge and because the CDC Influenza Division recommends its usage, many other laboratories worldwide will be using InGen. In addition there are basic users who can choose to share data through groups, but then there are managers of influenza (Table 3.1). These managers will be picked by the CDC Influenza Division, and usage of InGen gives rights to these particular users for surveillance of influenza. With the combination of managers and worldwide usage, InGen will have the capability to remove any significant lag time between local laboratories and national laboratories for influenza surveillance.

Table 3.1. InGen user classes.

	Access level	Magnitude of estimated users
Administrator	Have total control over InGen and are likely to be the developers of the project. Administrators can delete users, view information about users, and log in as another user.	ones
Manager	May view all data as it pertains to one organism (especially influenza), without regard to which group the data belong to.	tens
Group Administrator	Each group administrator is in control of a particular group. May add or remove users from a group. May alter permissions of group members as it relates to the group (view, edit, etc.).	hundreds
User	May view and edit their set data. May view and edit group data, as allowed by each group administrator. May create groups and consequently become group administrators of those groups.	thousands

CHAPTER 4

CG-PIPELINE, A COMPUTATIONAL GENOMICS PIPELINE FOR PROKARYOTIC SEQUENCING PROJECTS

ABSTRACT

Motivation: New sequencing technologies have accelerated research on prokaryotic genomes and have made genome sequencing operations outside major genome sequencing centers routine. However, no off-the-shelf solution exists for the combined assembly, gene prediction, genome annotation, and data presentation necessary to interpret sequencing data. The resulting requirement to invest significant resources into custom informatics support for genome sequencing projects remains a major impediment to the accessibility of high-throughput sequence data.

Results: We present CG-Pipeline, a self-contained, automated, high-throughput, open source genome sequencing and computational genomics pipeline suitable for prokaryotic sequencing projects. The pipeline has been used at the Georgia Institute of Technology and the Centers for Disease Control and Prevention for the analysis of *Neisseria meningitidis* and *Bordetella bronchiseptica* genomes. The pipeline is capable of enhanced or manually assisted reference-based assembly using multiple assemblers and modes; gene predictor combining; and functional annotation of genes and gene products. Because every component of the pipeline is executed on a local machine with no need to access resources over the Internet, the pipeline is suitable for projects of a sensitive nature. Annotation of virulence-related features makes the pipeline particularly useful for projects working with pathogenic prokaryotes.

Availability and Implementation: The pipeline is licensed under the open-source GNU General Public License and available at the Georgia Tech *Neisseria* Base (http://nbase.biology.gatech.edu/). The pipeline is implemented with a combination of Perl, Bourne Shell, and MySQL and is compatible with Linux and other Unix systems.

INTRODUCTION

Genome sequencing projects, pioneered in the 1990s (Fleischmann *et al.*, 1995), require large-scale computational support in order to make their data accessible for use and interpretation by biologists. Large sequencing centers have traditionally employed or collaborated with teams of software engineers and computational biologists to develop the software and algorithms for sequencing hardware interfaces, enterprise data storage, sequence assembly and finishing, genome feature prediction and annotation, database mining, comparative analysis, and database user interface development. While many of the components developed by these teams are now available online under open-access terms, the development of new, high-throughput sequencing technologies has necessitated updates to these tools and development of even more sophisticated algorithms to address the challenges raised by the new data. These new technologies – 454 pyrosequencing (Margulies *et al.*, 2005), ABI SOLiD (Shendure *et al.*, 2005), and Illumina (Bentley *et al.*, 2008) – are now collectively referred to as second generation sequencing technologies. Similar updates will be needed as the third generation of sequencing technologies, such as Pacific Biosciences' SMRT sequencing (Eid *et al.*, 2009), enter production use. New and improved tools released for these technologies on a monthly basis include assemblers, mapping algorithms, base calling and error correction tools, and a multitude of other programs. Because of this fast pace of development, few experts are able to keep up with the state of the art in the field of computational genomics. Accordingly, the

rate limiting step in genome sequencing projects is no longer the experimental characterization of the data but rather the availability of experts and resources for computational analysis.

At the same time, the increased affordability of these new sequencing machines has spawned a new generation of users who were previously unable to perform their own genome sequencing, and thus collaborated with large sequencing centers for genome sequencing and subsequent computational analysis. While these users are now able to characterize genomes experimentally in house, they often find themselves struggling to take full advantage of the resulting data and to make it useful to the scientific community since the informatics support for their genome projects is not sufficient.

Several large sequencing consortia (Aziz *et al.*, 2008; Markowitz *et al.*, 2009; Seshadri *et al.*, 2007) have produced comprehensive, centralized web-based portals for the analysis of genomic and metagenomic data. While extremely useful for many types of projects and collaborations, these solutions inherently result in a loss of data processing flexibility compared to locally installed resources and may be unsuitable for projects dealing with sensitive data. Recently, another group (Stewart *et al.*, 2009) has published DIYA, a software package for gene prediction and annotation in bacterial genomes with a modularized, open source microbial genome processing pipeline. However, DIYA does not include a genome assembly component, and does not provide for the combination of complementary algorithms for genome analysis.

To address the outstanding challenges for local computational genomics support, we have developed a state of the art, self-contained, automated high-throughput open source software pipeline for computational genomics in support of prokaryotic sequencing projects, which we call CG-Pipeline, for "Computational Genomics Pipeline." To ensure the relevance of our pipeline, we checked the latest developments in computational genomics software for all stages of the pipeline, such as new versions of assembly and gene prediction programs and

comparative surveys, and selected what we deemed to be the most suitable software packages. The pipeline is self-contained; that is, we used locally installable versions of all third-party tools instead of web-based services provided by many groups. We chose to do so for three reasons: first, because some of the applications we envision for this pipeline are of sensitive nature; second, to enhance robustness to external changes (*e.g.*, online API changes or website address changes); and third, to improve the ability of developers to customize and derive from our pipeline. The pipeline is also automated and high-throughput: all components are organized in a hierarchical set of readily modifiable scripts, and the use of safe programming practices ensures that multiple copies of the pipeline can be run in parallel, taking advantage of multiple processors where possible.

Importantly, by using and combining the outputs of competitive, complementary algorithms for multiple stages of genome analysis, our pipeline allows for substantial improvement upon single-program solutions. The use of multiple algorithms also provides a way to improve robustness and conduct more comprehensive quality control when the output of one program is significantly different from that of another.

Computational support provided to prokaryotic genome projects by our pipeline can be subdivided into three stages: first, sequencing and assembly; second, feature prediction; and third, functional annotation. For the assembly stage, we developed a custom protocol specific to 454 pyrosequenced data, which resulted in a significant improvement to assembly quality of our test data compared to the baseline assembler bundled by the manufacturer. Other assemblers can be plugged in if necessary, and data from other sequencing technologies such as ABI SOLiD, Illumina and Sanger capillary-based machines can be used. For the prediction stage, we again included a custom combination of feature prediction methods for protein-coding genes, RNA genes, operon and promoter regions, which improves upon the individual constituent methods.

The annotation stage includes several types of protein functional prediction algorithms. We also developed components for comparative analysis, interpretation and presentation (a web-based genome browser), which can be used downstream of our pipeline.

We have tested the pipeline on the bacterium *Neisseria meningitidis*, which is a human commensal of the nasopharynx and which can sometimes cause meningitis or septicemia (Rosenstein *et al.*, 2001). When *N. meningitidis* does cause disease, it can be devastating with an approximately 10% fatality rate and 15% sequelae rate. *N. meningitidis* is a highly competent organism with a high recombination rate, and large chromosomal changes are common (Jolley *et al.*, 2005; Schoen *et al.*, 2008). This complicates computational genome analysis and makes *N. meningitidis* an appropriately challenging test for our pipeline. To demonstrate the general applicability of the pipeline, we have also tested it on a different pathogen, *Bordetella bronchiseptica*. *B. bronchiseptica* is a Gram-negative bacterium that can cause bronchitis in humans, although it is more commonly found in smaller mammals (Parkhill *et al.*, 2003). Much like *Neisseria*, *Bordetella* has extensive plasticity, likely due to the large number of repeat elements (Gerlach *et al.*, 2001). Here, we analyze the first two complete genome sequences of *B. bronchiseptica* strains isolated from human hosts.

The rest of this paper is organized as follows. The System and Methods section describes the genomes which we used to test CG-Pipeline, overall organization of the pipeline, and details of the algorithms used to perform tasks in the pipeline. In the Discussion section, we discuss the objectives of our work on the pipeline and how these relate to larger developments in computational biology for next-generation sequencing.

SYSTEM AND METHODS

GENOME TEST DATA

N. meningitidis genomes were characterized via 454 pyrosequencing (Margulies *et al.*, 2005) using either a half or one quarter plate runs on the Roche 454 GS-20 or GS Titanium instrument (Table 4.1). For each genome, a random shotgun library was produced using Roche protocols for nebulization, end-polishing, adaptor ligation, nick repair and single-stranded library formation. Following emulsion PCR, DNA bound beads were isolated and sequenced using long read (LR) sequencing kits. The number of reads produced in the experiments ranged from 200,000 to 600,000, and the average read lengths were between 100 and 330 bases. These data yielded 47.6-94.3 million bases per genome amounting to 20-40x coverage for the approx. 2.1 megabase *N. meningitidis* genomes. After read trimming and re-filtering to recover short quality reads, the data were passed to the first stage of the pipeline – genome assembly.

CG-PIPELINE DEVELOPMENT

The CG-Pipeline has been developed in two phases. The first fully functional version of CG-Pipeline (v0.2) was released in June 2010 (Kislyuk et al., 2010). The second phase of development (CG-Pipeline v0.3) entailed a number of enhancements in response to user requests and new data sources. Below, we describe the distinct functionality and features made available for each stage of analysis in CG-Pipeline versions 0.2 and 0.3.

Table 4.1. Summary of sequencing projects used in the pipeline development.

Strain ID	ST/SG [a]	Origin [b]	Genome size	Closest Reference [c]	Total reads	Total bases sequenced	Avg read length	Coverage [d]	454 standard [e]
Neisseria meningitidis									
M13220	7/A	Philippines 2005	2.2M	Z2491	197067	47569493	241	21×	GS-20
M10699	32/B	Oregon, USA 2003	2.2M	MC58	418751	81775264	195	37×	GS-20
M15141	11/C	New York, USA 2006	2.2M	FAM18	378773	94288660	249	42×	GS-20
M9261	11/W135	Burkina Faso 2002	2.2M	FAM18	206634	69957473	338	31×	GS Ti
M18575	2859/A	Burkina Faso 2003	2.2M	Z2491	283888	84013571	296	38×	GS Ti
M5178	32/B	Oregon, USA 1998	2.2M	MC58	270332	88664981	328	40×	GS Ti
M15293	32/B	Georgia, USA 2006	2.2M	MC58	276733	90951566	329	41×	GS Ti
Bordetella bronchiseptica									
BBE001	N/A [f]	Georgia, USA 1956	5.3M	RB50	566834	229098141	404	43×	GS Ti
BBF579	N/A	Mississippi, USA 2007	5.3M	RB50	533099	228467710	429	43×	GS Ti

[a] Sequence type (ST) denotes the allelic profile assigned by multilocus sequence typing (MLST) (Holmes *et al.*, 1999; Maiden *et al.*, 1998) on the basis of seven loci within well-conserved house-keeping genes. *N. meningitidis* isolates are divided into serogroups (SG) by immunochemistry of polysaccharides present in their antiphagocytic capsule.

[b] The geographic region in which each strain was originally collected and the date that the isolate was collected.

[c] Strain ID of the closest complete genome available in GenBank, as determined by 16S RNA phylogeny as well as whole-genome sequence identity, which agreed in all cases.

[d] Coverage denotes the average number of sequencing reads overlapping at a given position in the genome, calculated as the total number of bases sequenced divided by the estimated length of the genome.

[e] The standard of the 454 pyrosequencing instrument and reagents used to sequence the data.

[f] Sequence typing and serotyping was not performed on *B. bronchiseptica*.

PIPELINE ORGANIZATION V0.2

The analytical pipeline consists of three integrated stages: genome assembly, feature prediction and functional annotation. Each stage consists of a top-level execution script managing the input, output, format conversion, and combination of results for a number of distinct software components. A hierarchy of scripts and external programs then performs the tasks required to complete each stage of analysis (Figure 4.1)

PIPELINE ORGANIZATION V0.3

The overall architecture of CG-Pipeline has been retained for version 0.3 and additional features have been added to each of the three stages (Figure 4.2). The assembly stage now includes an additional assembly software component, the Velvet assembler, along with an automated tool for comparing multiple assemblies and selecting the best assembly based on standard annotation metrics. The feature prediction stage has been enhanced by the addition of the RNAmmer software component that predicts ribosomal RNA loci. The functional annotation stage has been modified to produce three standard output files containing the sequences and locations of all annotated features. These three distinct file formats – GenBank, GFF, FASTA – are all standard and portable among multiple bioinformatics applications. Finally, the new version includes a single wrapper script around all three stages that allows for fully integrated automation of the entire pipeline. This enhancement moves the pipeline closer to its ultimate goal of fully functional single click genome analysis.

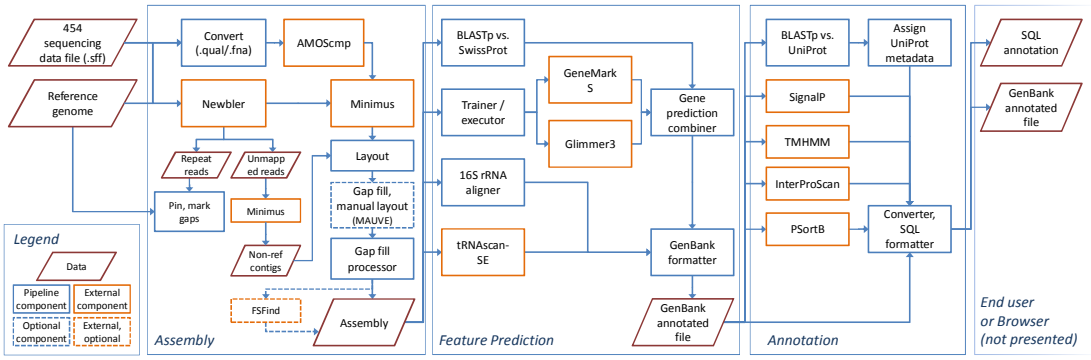


Figure 4.1. Chart of data flow, major components and subsystems in the CG-Pipeline v0.2. Three subsystems are presented: genome assembly, feature prediction and functional annotation. Each subsystem consists of a top-level execution script managing the input, output, format conversion, and combination of results for a number of components. A hierarchy of scripts and external programs then performs the tasks required to complete each stage. The legend for the flowchart indicates the identities of the distinct pipeline components: data, pipeline component, optional component, external component and external, optional component.

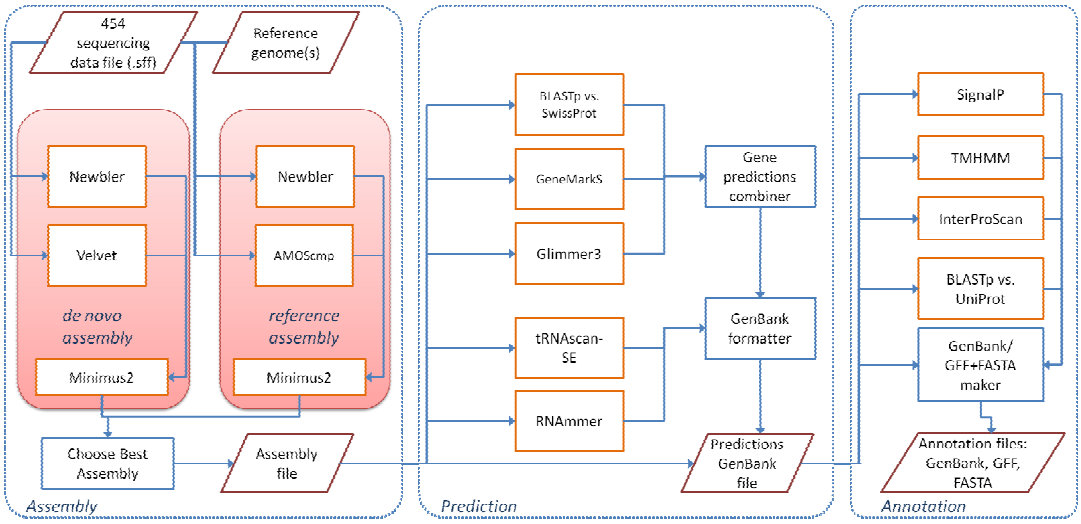


Figure 4.2. Chart of data flow, major components and subsystems in CG-Pipeline v0.3. Dotted lines indicate the three subsystem components: genome assembly, feature prediction and functional annotation. Red parallelograms indicate input/output files, orange boxes show individual program components, and blue boxes indicate custom scripts. Distinct subassembly pipelines discussed in the text are highlighted in pink.

ASSEMBLY V0.2

Genome assembly was performed by evaluating multiple configurations of assemblers including the standard 454 assembler, Newbler (version 2.3), as well as the Celera Assembler (Miller *et al.*, 2008), the Phrap assembler (http://www.phrap.org/) and the AMOScmp mapped assembler (Pop *et al.*, 2004). Several other assemblers were evaluated but ultimately excluded from the pipeline due to use limitations: for instance, the ALLPATHS 2 assembler (MacCallum *et al.*, 2009) required paired-end reads to operate; our evaluation data contained no paired-end reads, and such a requirement unnecessarily constrains the user's options. The widely used Velvet assembler (Zerbino & Birney, 2008) was originally developed as a *de novo* assembler for Illumina sequencing technology, but its capability has been extended to accommodate 454 data as well. However, we were unable to configure the Velvet assembler to produce a usable assembly or take advantage of reference genomes using 454 data alone.

Comparative evaluation of our initial assembly results indicated that reference assemblies of *N. meningitidis* genomes using previously finished strains were of superior quality to *de novo* assemblies. Using the most appropriate reference strains, it was found that Newbler and AMOScmp complement each other's performance in the reference assembly stage, with Newbler being able to join some contigs AMOScmp left gapped and *vice versa*. As a result, we decided to use a combination of these two assemblers' outputs for the final assembly. Then, the Minimus assembler (Sommer *et al.*, 2007) from the AMOS package, a simple assembler for short genomes, was used to combine the constituent assemblies.

We also evaluated alternative base calling algorithms for 454 pyrosequencing data (Quinlan *et al.*, 2008) but detected no improvement. Over the course of our project, accuracy of base calling in the Newbler assembler was reported to be significantly improved. We used the latest version of the assembler available at publication time (v2.3).

An optional component of the pipeline was created for frameshift detection using FSFind (Kislyuk *et al.*, 2009). Frameshifts in protein-coding sequences are a known result of pyrosequencing errors caused by undercalls and overcalls in homopolymer runs (Kuo & Grigoriev, 2009). Briefly, this package creates a GeneMark model of the genome, makes gene predictions, and then scans the genome for possible frameshift positions on the basis of ORF configuration and coding potential. Once the possible frameshift sites are identified, a putative translation of the protein possibly encoded by the broken gene is compared against a protein database (SwissProt by default). The predicted frameshift site is also scanned for adjacent homopolymers. A heuristic set of confidence score cutoffs is then used to provide a set of frameshift predictions while minimizing the false positive rate. The predicted frameshift sites can then be verified experimentally or corrected speculatively. The user can inspect the dataset to decide whether locations predicted to contain frameshifts break gene models, and patch the sequences to fix up these positions. The prediction stage can then be re-run to correct the gene predictions. While further experimental analysis to address such errors is desirable (*e.g.*, targeted PCR of predicted error locations or a recently popular choice of combining sequencing technologies such as 454 and Illumina), it incurs extra costs which we aim to avoid.

Unfinished assemblies produced in this stage contained 90-300 contigs each. No paired-end libraries or runs were available for the strains analyzed, and therefore scaffolding of the contigs was a challenge. Manual examination of the assemblies using the MAUVE (Darling *et al.*, 2004) multiple whole-genome alignment and visualization package revealed numerous locations where contigs could be scaffolded with a small gap or minimal overlap (Figure 4.3). As an optional step, we produced a table of such positions and a script which would scaffold contigs joined by the gap. Then, a manual gap joining stage used the layout of the contigs according to their aligned positions on the reference using the AMOS package and manual examination of

each gap, adjacent contig alignments and reference annotation in the MAUVE visualization tool. Although there is a possibility that rearrangements exist in those gaps as mapped to the closest reference genome, joining was only done after manual examination on a case-by-case basis in positions of high homology and full consensus between four of the reference strains, to minimize this possibility. While we provide the scripts and data format definitions necessary to complete this stage of the pipeline, it involves manual processing of the assembly and is therefore optional. This component is similar in function to Mauve Contig Mover (Rissman *et al.*, 2009) but expands upon it in several ways. An option is provided in the pipeline to use Mauve Contig Mover.

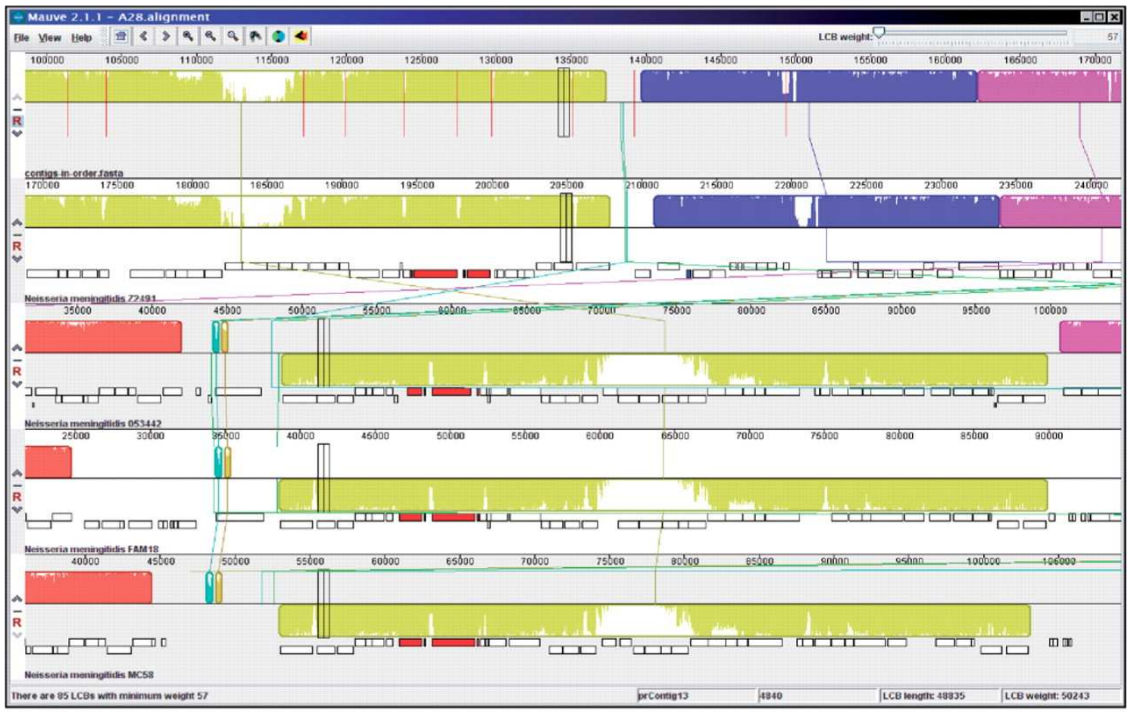


Figure 4.3. Comparative analysis of draft assembly with MAUVE. The top pane represents the active assembly; vertical lines indicate contig boundaries (gaps). The reference genomes are arranged in subsequent panes in order of phylogenetic distance. Blocks of synteny (LCBs) are displayed in different colors (an inversion of a large block is visible between panes 1–2 and 3–5). Most gaps within LCBs were joined in the manually assisted assembly, while considering factors such as sequence conservation on contig flanks and presence of protein-coding regions.

The manually assisted genome assembly procedure resulted in an order-of-magnitude decrease in the number of gaps in comparison to the Newbler assembler (which in turn performed the best out of all standalone assemblers evaluated). In addition, the fully automated assembly metrics (N50 and contig count at equal minimal size) are an approximately 20-50% improvement upon baseline Newbler performance (Table 4.2).

Table 4.2. Summary of assembler performance (v0.2). Data for each strain are presented in rows. Statistics from standalone assemblers (Newbler and AMOScmp) are presented together with results of the combining protocol (default output of the pipeline) and an optional, manually assisted predictive gap closure protocol.

Strain ID	**Newbler statistics**		**AMOScmp statistics**	
	Contigs > 500 nt, total size	N50[a], Longest contig	Contigs > 500 nt, total size	N50, Longest contig
M13220	175, 2.07M	22K, 106K	202, 2.06M	21K, 77K
M10699	102, 2.10M	52K, 143K	116, 2.10M	43K, 113K
M15141	147, 2.06M	33K, 171K	190, 2.05M	22K, 115K
M9261	99, 2.09M	51K, 184K	133, 2.07M	37K, 170K
M18575	133, 2.09M	30K, 172K	147, 2.09M	29K, 88K
M5178	89, 2.13M	56K, 136K	107, 2.12M	42K, 131K
M15293	92,2.08M	52K, 144K	110, 2.06M	42K, 132K
BBE001	146, 5.05M	70K, 212K	178, 5.04M	61K, 173K
BBF579	272, 4.84M	57K, 88K	321, 4.84M	46K, 94K
	Automatic combined assembly		**Manual combined assembly**	
	Contigs > 500 nt, total size	N50, Longest contig	Contigs > 500 nt, total size	% gapfill, Longest contig
M13220	195, 2.25M	31K, 107K	57, 2.30M	1.8%, 398K
M10699	83, 2.17M	59K, 143K	40, 2.18M	1.1%, 435K
M15141	139, 2.21M	36K, 171K	50, 2.28M	2.0%, 759K
M9261	128, 2.16M	64K, 231K	27, 2.21M	1.6%, 866K
M18575	220, 2.40M	53K, 231K	N/A[c]	N/A
M5178	104, 2.17M	59K, 136K	N/A	N/A
M15293	107, 2.10M	59K, 144K	N/A	N/A
BBE001	214, 5.03M	80K, 252K	N/A	N/A
BBF579	272, 4.84M[b]	57K, 88K	N/A	N/A

[a] N50 is a standard quality metric for genome assemblies that summarizes the length distribution of contigs. It represents the size N such that 50% of the genome is contained in contigs of size N or greater. Greater N50 values indicate higher quality assemblies.

[b] No improvement was detected from the combined assembly in strain BBF579, and the original Newbler assembly was automatically selected. (c) The manual combined assembly protocol was not performed for these projects.

Table 4.3. Summary of assembler performance (v0.3). Data for each strain are presented together with results of the combining protocol (default output of the pipeline).

Strain ID	Automated combined assembly			
	Contigs > 500 nt	Total size	N50[a]	Longest contig
M13220	168	2.17M	28K	80K
M10699	87	2.20M	52K	143K
M15141	124	2.05M	34K	233K
M9261	73	2.12M	69K	230K
M18575	87	2.37M	45K	172K
M5178	76	2.30M	56K	136K
M15293	81	2.16M	56K	144K
BBE001	126	5.09M	80K	212K
BBF579	224	5.08M	41K	96K

[a] N50 is a standard quality metric for genome assemblies that summarizes the length distribution of contigs. It represents the size N such that 50% of the genome is contained in contigs of size N or greater. Greater N50 values indicate higher quality assemblies.

The contigs in the assembly stage output were named according to the following format: `prefix_contig#`, where the prefix represents a unique strain identifier and # represents the zero-padded sequential number indicating the contig's predicted order on the chromosome. For example, the 25th contig for the *N. meningitidis* strain M13220 assembly would be named as `CDC_NME_M13320_025`. The prefix used in the pipeline is configurable by the user with a command line option.

ASSEMBLY V0.3

I have incorporated the Velvet assembler at the assembly stage (Figure 4.2). As noted in the original CG-Pipeline publication Velvet does not add much to the final output of CG-Pipeline (Kislyuk et al., 2010). However, it has bettered the final metrics somewhat with about a 2-5% reduction in the final number of contigs (data not shown). The main reason that Velvet has been incorporated is that CG-Pipeline has been requested by users to be applied to Illumina sequence data. In the future CG-Pipeline will be ready to be used on Illumina data.

The assembly stage has been modified at the combining stage. For any assembly in CG-Pipeline multiple assemblers may produce an assembly, and those assemblies are combined into a comprehensive assembly. Different reference assemblies are not combined, nor are de novo with reference. I have replaced Minimus with Minimus2 (Sommer et al., 2007). We chose Minimus originally because it is a good short read assembler. However, users encountered fatal problems due to its large memory usage. We experimented with Minimus2 and found better memory usage i.e. no fatal problems due to memory usage, faster processing, and better resulting assemblies (unpublished data). Fortunately, the incorporation of Minimus2 required very little change to the assembly pipeline code and was easily accomplished.

The assembly stage has also been modified such that the best comprehensive assembly is chosen before continuing to feature prediction. In this new paradigm, a user may choose multiple reference genomes for assembly. The assembly stage will create an assembly against each reference genome in addition to a *de novo* assembly. Thus the assembly stage will have multiple candidates as the best assembly: one *de novo* and one reference assembly per reference. The best assembly is picked based on several standard assembly metrics: number of contigs, total bases assembled, average contig size, and N50. N50 is defined as a size N such that 50% of the genome is contained in contigs of size N or greater. This step of producing many assemblies and choosing of the best assembly is automated and does not require human intervention.

FEATURE PREDICTION V0.2

Feature prediction was performed in the genome using a suite of several programs. To predict genes, we used a combination of *de novo* and comparative methods. The Glimmer (Delcher *et al.*, 1999) and GeneMark (Besemer *et al.*, 2001) microbial gene predictors were used for *de novo* prediction, and BLASTp alignment (Altschul *et al.*, 1997) of putative proteins was

used for comparative prediction. Self-training procedures were followed for both *de novo* predictors, and the results, while highly concordant, were different enough (Table 4.4) to justify the inclusion of both algorithms. BLASTp alignment of all open reading frames (ORFs) at least 90 nt long was performed using the Swiss-Prot protein database (Boeckmann *et al.*, 2003).

Table 4.4. Prediction algorithm performance comparison and statistics (v0.2). Data for each strain are presented in rows. Prediction counts from the 3 standalone gene prediction methods are presented. Counts of protein-coding gene predictions reported by our algorithm and tRNA genes are also shown. Data presented are based on the automatic combined assemblies from Table 2.

Strain ID	Gene predictions by GeneMark	Gene predictions by Glimmer3	Gene predictions by BLAST	ORFs with full consensus[a]	ORFs with partial consensus[b]	Total gene predictions reported[c]	tRNAs predicted by tRNAScan-SE
M13220	2530	2725	1353	1325	974	2299	52
M10699	2366	2494	1317	1284	826	2110	51
M15141	2411	2578	1369	1343	841	2184	57
M9261	2370	2553	1341	1308	802	2110	51
M18575	2751	2927	1495	1448	1023	2471	63
M5178	2377	2510	1315	1281	816	2097	52
M15293	2062	2040	1285	1261	802	2063	51
BBE001	4793	4793	2744	2732	2067	4799	48
BBF579	4649	4646	2652	2635	2021	4656	48

[a] Number of ORFs with protein-coding gene predictions where all 3 predictors agreed exactly or with a slight difference in the predicted start site.

[b] ORFs where only 2 of the 3 predictors made a prediction.

[c] Total protein-coding gene predictions reported by the pipeline.

Table 4.5. Prediction algorithm performance comparison and statistics (v0.3). Data for each strain are presented in rows. Counts of protein-coding gene predictions, rRNA genes, and tRNA genes are shown.

Strain ID	Total gene predictions reported	tRNAs predicted by tRNAScan-SE	rRNAs predicted by RNAmmer	Protein-coding genes
M13220	2225	51	0	2174
M10699	2194	51	0	2143
M15141	2088	53	0	2035
M9261	2134	51	0	2083
M18575	2487	59	0	2428
M5178	2300	54	1	2245
M15293	2180	55	1	2124
BBE001	4907	47	1	4859
BBF579	4888	47	1	4840

The results of these three methods were combined together using a combiner strategy outlined in Figure 4.4. In this strategy, we first check that at least half of the predictors report a gene in a given ORF – in our configuration, 2 of the 3 predictors. Then the Met (putative translation start) codon closest to the beginning of the BLAST alignment is found and declared to be the gene start predicted by BLAST. We then find the gene start coordinate reported by the majority of the three predictors and report the resulting gene prediction. If no majority exists, we select the most upstream gene start predicted.

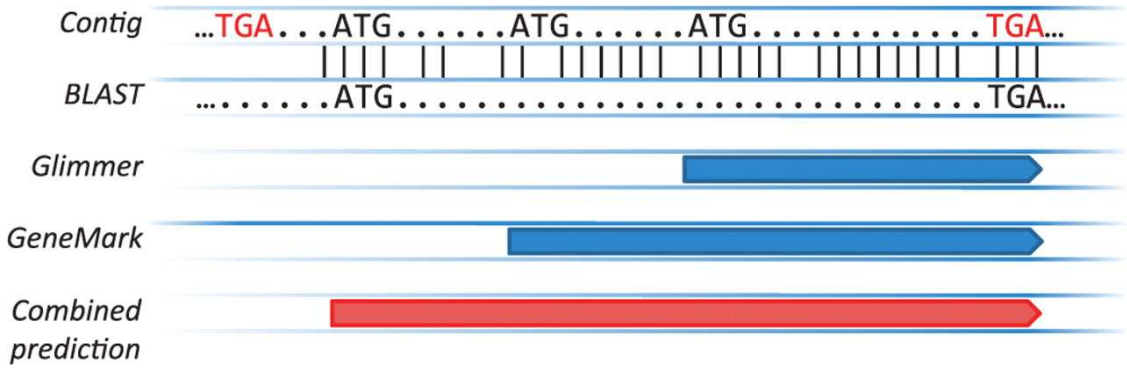


Figure 4.4. Schematics of combining strategy for prediction stage. BLAST alignment start, which may not coincide exactly with a start codon, is pinned to the closest start codon. Then, a consensus or most upstream start is selected.

In addition to protein-coding gene prediction, ribosomal genes were predicted using alignment to a reference database of ribosomal operons, and tRNA genes were predicted using the tRNAScan-SE package (Lowe & Eddy, 1997). The results are summarized in Table 4.4.

Results of the feature prediction stage are saved in a multi-extent GenBank formatted file. Features were named according to the following convention: `contig-name_feature-id`, where `contig-name` is as described earlier, and `feature-id` is a sequential zero-padded number unique to the feature across all contigs. For example, a gene with feature ID 1293 on contig 25 might have the name `CDC_NME_M13320_025_1293`.

To validate the overall accuracy of the gene prediction stage of the pipeline, we ran our gene prediction tools on the genome of *Escherichia coli* K12*,* one of the best-annotated bacterial genomes (analysis described in Appendix B). Our pipeline was able to detect 95.7% of the annotated *E. coli* K12 protein-coding genes, and exactly predict starts in 85.5% of those. 50% of the E. coli predictions that report incorrect start codons report starts within 35 nt of the true start, and all reported starts are within 200 nt of the true start.

FEATURE PREDICTION V0.3

The feature prediction stage in CG-Pipeline is very robust. However, we were able to increase its capabilities by incorporating rRNA prediction. Here, we used RNAmmer (Lagesen *et al.*, 2007). RNAmmer uses hidden markov models to detect regions that likely belong to rRNA loci. RNAmmer has been able to detect rRNAs with 100% specificity and sensitivity in four reference meningococcal genomes (data not shown). With RNAmmer and the other feature predictors used in the pipeline (Altschul *et al.*, 1997; Apweiler *et al.*, 2010; Delcher *et al.*, 1999; Lowe & Eddy, 1997; Lukashin & Borodovsky, 1998), feature prediction is able to detect >95% of all features (Kislyuk *et al.*, 2010). Updated feature prediction metrics are given in Table 4.5.

FUNCTIONAL ANNOTATION V0.2

Functional annotation of genome features was also performed using a combination of tools. Annotation of protein coding genes was based on an integrated platform that makes use of six distinct annotation tools, four of which employ intrinsic sequence characteristics for annotation and two that use extrinsic homology-based approaches to compare sequences against databases of sequences and structures with known functions. Information on Gene Ontology (GO) terms, domain architecture and identity, subcellular localization, signal peptides, transmembrane helices and lipoprotein motifs is provided for each protein coding gene (Figure 4.5).

BLASTp alignment of predicted proteins was performed against the UniProt database (Uniprot, 2009). Homology-based searches were also made across thirteen sequence and protein domain databases with the InterProScan suite (Mulder & Apweiler, 2007). Parsing of the results was carried out against the corresponding InterPro database. The pipeline also stores the top five hits for each gene against the NCBI non-redundant protein database, to provide potentially useful information. All homology searches were run locally. Signal peptides were annotated using the SignalP package (Bendtsen *et al.*, 2004) and transmembrane domains were annotated with the TMHMM package (Krogh *et al.*, 2001). State of the art in subcellular localization algorithms was examined to ensure the best performance given our operational requirements. Insertion sequences (transposases) and proteins reported as virulence factors by VFDB (Chen *et al.*, 2005; Yang *et al.*, 2008) were also annotated. These annotations of virulence-related features make the pipeline particularly useful for projects working with pathogenic prokaryotes. Results of this analysis are summarized in Table 4.6.

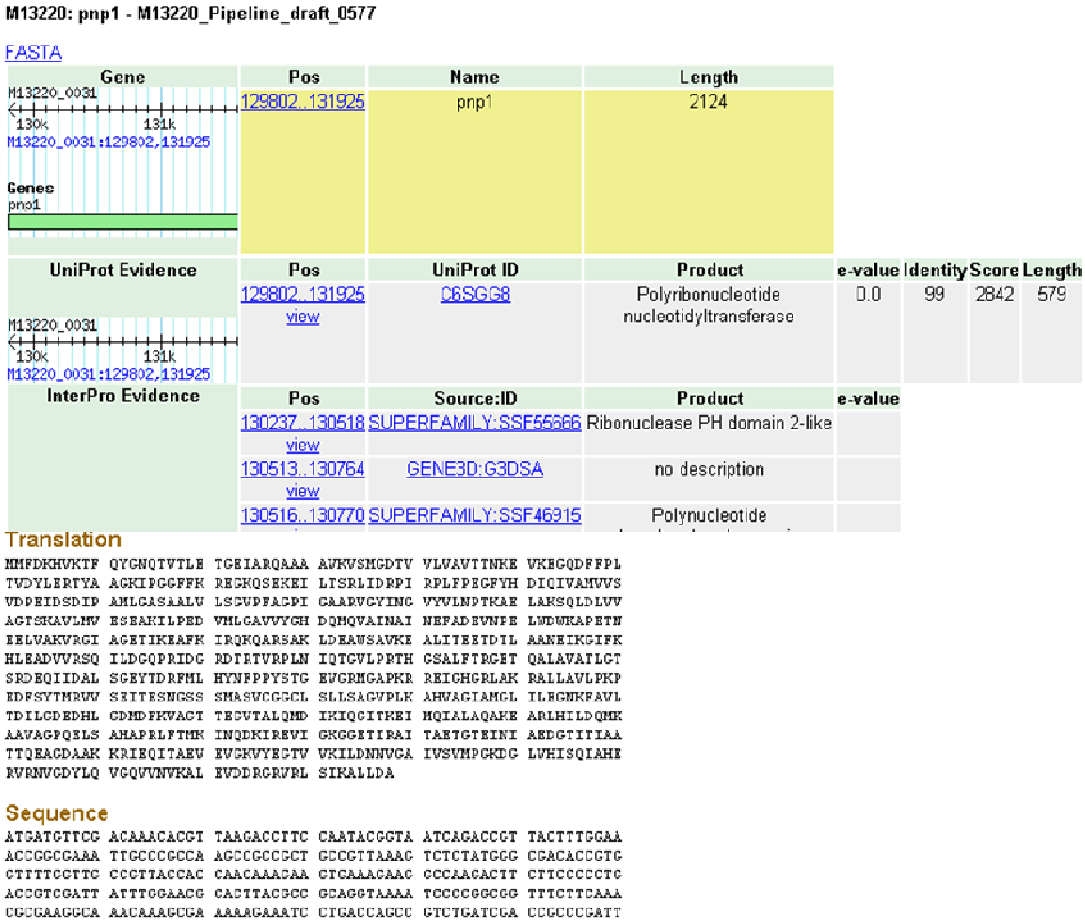


Figure 4.5. Example functional annotation listing of an *N.meningitidis* gene in *Neisseria* Base. Draft genome data are shown including gene location, prediction and annotation status, peptide statistics, BLAST hits, signal peptide properties, transmembrane helix presence, DNA and protein sequence. All names, locations, functional annotations and other fields are searchable, and gene data are accessible from GBrowse genome browser tracks.

Table 4.6. Feature annotation statistics (v0.2). Data for each strain are presented in rows. Data presented are based on the automatic combined assemblies from Table 4.2 and the gene predictions from Table 4.4.

Strain ID	Total number of CDS[a]	Signal peptides[b]	Transmembrane helices[c]	Conserved hypothetical proteins	Putative uncharacterized proteins	Functional assignment inferred from homology	Virulence factors[d]
M13220	2299	326 (14.2%)	184 (8.0%)	10 (0.4%)	708 (30.8%)	603 (26.2%)	36 (1.6%)
M10699	2110	310 (14.7%)	180 (8.5%)	5 (0.2%)	652 (30.9%)	577 (27.3%)	45 (2.1%)
M15141	2184	317 (14.5%)	173 (7.9%)	16 (0.7%)	590 (27.0%)	583 (26.7%)	50 (2.3%)
M9261	2110	303 (14.4%)	166 (7.9%)	13 (0.6%)	591 (28.0%)	558 (26.4%)	37 (1.8%)
M18575	2471	349 (14.1%)	193 (7.8%)	13 (0.5%)	725 (29.3%)	668 (27.0%)	48 (1.9%)
M5178	2097	298 (14.2%)	177 (8.4%)	3 (0.1%)	646 (30.8%)	572 (27.3%)	45 (2.1%)
M15293	2063	304 (14.7%)	168 (8.1%)	6 (0.3%)	613 (29.7%)	567 (27.5%)	47 (2.3%)
BBE001	4799	977 (20.4%)	368 (7.7%)	9 (0.2%)	807 (16.8%)	1184 (24.7%)	54 (1.1%)
BBF579	4656	934 (20.1%)	339 (7.3%)	9 (0.2%)	739 (15.9%)	1171 (25.2%)	45 (1.0%)

[a] Total putative protein-coding sequences analyzed.

[b] As predicted by SignalP (Bendtsen *et al.*, 2004) ; percentage of total CDS indicated in parentheses.

[c] As predicted by TMHMM (Krogh *et al.*, 2001).

[d] As predicted by BLASTp alignment against VFDB (Chen *et al.*, 2005; Yang *et al.*, 2008) ; http://www.mgc.ac.cn/VFs/

Table 4.7. Feature annotation statistics (v0.3). Data for each strain are presented in rows. Annotations are predicted upon the predictions in Table 4.5, which predicted features from the assemblies in Table 4.3.

Strain ID	Total number of CDS[a]	Signal peptides[b]	Transmembrane helices[c]	Functional assignment inferred from homology	Virulence factors[d]
M13220	2174	268 (12.2%)	406 (18.5%)	730 (33.3%)	33 (1.5%)
M10699	2143	255 (11.9%)	388 (18.1%)	679 (31.7%)	45 (2.1%)
M15141	2035	243 (11.9%)	378 (18.6%)	673 (33.1%)	46 (2.3%)
M9261	2083	256 (12.3%)	385 (18.5%)	684 (32.8%)	37 (1.8%)
M18575	2428	276 (11.4%)	442 (18.2%)	788 (32.5%)	41 (1.7%)
M5178	2245				
M15293	2124	251 (11.8%)	406 (19.1%)	680 (32.0%)	48 (2.3%)
BBE001	4859				
BBF579	4840	868 (17.7%)	1015 (20.0%)	987 (20.4%)	45 (0.9%)

[a] Total putative protein-coding sequences analyzed.

[b] As predicted by SignalP (Bendtsen *et al.*, 2004) ; percentage of total CDS indicated in parentheses.

[c] As predicted by TMHMM (Krogh *et al.*, 2001).

[d] As predicted by BLASTp alignment against VFDB (Chen *et al.*, 2005; Yang *et al.*, 2008) ; http://www.mgc.ac.cn/VFs/

After the functional annotations were determined, a naming scheme was employed for each locus to conform to standard annotation terminology. Specific gene names were assigned according to homology-based results. For genes that had a Uniprot result with a best hit at greater than 91% amino acid sequence identity and an e-value less than 1e-9, the gene assumed the best hit's name. If the best hit had the keyword "hypothetical," then we used a domain name from InterPro to name the gene. For example, if a gene was given the name "hypothetical" from Uniprot and a domain name of "transferase" from InterPro, then the final name was "hypothetical transferase protein." Therefore most genes that were given "hypothetical" or "putative" prefixes could then be given a more comprehensive name based on further information such as domain names or protein functions. Genes with unknown functions found across many genomes were given the name "conserved hypothetical protein," and all

other putative genes with unknown functions were given the name "putative uncharacterized protein."

FUNCTIONAL ANNOTATION V0.3

Previously, CG-Pipeline outputted several SQL files that represented the annotation output. Unfortunately for casual users of CG-Pipeline, these files are nonstandard, meant to be parsed by a computer, and are difficult for a person to read. Therefore we added a tool to the annotation pipeline to compile a GenBank file from the output. The GenBank file includes information from all annotations performed by CG-Pipeline: BLAST against the Uniprot database (Altschul *et al.*, 1997; Apweiler *et al.*, 2010), InterProScan (Zdobnov & Apweiler, 2001), TMHMM (Krogh *et al.*, 2001), SignalP (Bendtsen *et al.*, 2004), BLAST against the Virulence Factors Database (Chen *et al.*, 2005; Yang *et al.*, 2008), and the noncoding features, namely rRNA and tRNA. The GenBank output file is comprehensive to all pipeline output and will include all future annotations that CG-Pipeline will make.

Fortunately because the GenBank format is standard, the output from CG-Pipeline is portable, meaning that it can be imported into any other bioinformatics application. For example, one potential use of the CG-Pipeline GenBank file is to upload it directly to a genome database such as NBase (Katz *et al.*, 2010) or GenBank (Benson *et al.*, 2010). Thus the time from isolating a strain to sequencing it to publishing a GenBank file could be on the order of 1-2 weeks. See Table 4.7 for updated annotation metrics.

AVAILABILITY

The pipeline software package is available at our website (http://nbase.biology.gatech.edu). The package contains detailed instructions and scripts for installation of the pipeline and all external programs, documentation on usage of the pipeline and its organization. Components

which require large biological databases automatically download local copies of those databases upon installation.

All of the *N. meningitidis* genomes reported here, along with custom annotations and tools for searching and comparative sequence analysis, are available for researchers online at our genome browser database (http://nbase.biology.gatech.edu).

DISCUSSION

GENOME BIOLOGY OF *N. MENINGITIDIS* AND *B. BRONCHISEPTICA*

We have used the pathogen *N. meningitidis* for the majority of developmental and production testing of our pipeline. Although *N. meningitidis* gains no fitness advantage from virulence, it occasionally leaves its commensal state and causes devastating disease (Meyers *et al.*, 2003). Several recent studies have used whole-genome analysis to determine the basis of virulence in this species but none have been conclusive (Hotopp *et al.*, 2006; Perrin *et al.*, 2002; Schoen *et al.*, 2008). With the recent advent of next-generation sequencing and the application of an analytical pipeline, such as presented here, this problem and other problems like it can be addressed in individual laboratories on a genome-wide scale. Here, we briefly speculate on a few of the implications of our findings for the genome biology of *N. meningitidis* to underscore the potential utility of our pipeline.

Whole genome analysis of microbes has led to the development of the 'pan-genome' concept (Tettelin *et al.*, 2005). A pan-genome refers to the collection of all genes found within different strains of the same species. An open pan-genome means that the genome of any given strain will contain unique genes not found within the genomes of other known strains of the same species. The extent to which microbial pan-genomes are open is a matter of debate (Lapierre & Gogarten, 2009). Recent studies have suggested that the *N. meningitidis* pan-genome is essentially open (Schoen *et al.*, 2008), consistent with the fact that it is known to be a

highly competent species (Chen & Dubnau, 2004; Kroll *et al.*, 1998). We evaluated this hypothesis by finding the number of unique genes in each of the seven strains reported here along with seven previously published strains, using the results of our analytical pipeline. Our findings are consistent with (Schoen *et al.*, 2008) in the sense that every genome sequence was found to contain at least 43 unique genes not found in any other strain. Thus, the *N. meningitidis* pan-genome does appear to be open.

N. meningitidis is a human commensal that most often does not cause disease, and avirulent strains of the species are referred to as carriage strains. Results of previous comparative genomic analyses have been taken to suggest that carriage strains represent a distinct evolutionary group that is basal to a group of related virulent strains of *N. meningitidis* (Schoen *et al.*, 2008). We tested this hypothesis using the results of our analytical pipeline applied to three carriage strains and eight virulent strains of *N. meningitidis*. Whole genome sequences were aligned and pairwise distances between genomes, based on nucleotide diversity levels, were compared within and between groups of carriage and virulent strains. We found that average of the pairwise genome sequence distances within (*w*) the carriage and virulent groups of strains was not significantly different from the average pairwise distances between (*b*) groups (w=0.074$\pm$0.027 b=0.090$\pm$0.014, t=0.693, P=0.491). This result is inconsistent with the previously held notion that carriage and virulent strains represent distinct evolutionary groups based on whole genome analysis. However, our findings are consistent with earlier work that found little genetic differentiation between carriage and virulent strains of *N. meningitidis* (Jolley, et al., 2005).

Currently, there is no unambiguous molecular assay to distinguish *B. bronchiseptica* from other *Bordetella* species. One reason the two *B. bronchiseptica* genomes reported here were characterized was to discover genes unique to the species (*i.e.* not present in any other

Bordetella species) to facilitate the development of a *B. bronchiseptica*-specific PCR assay. To identify such genes, we performed BLASTn with *B. bronchiseptica* query genes uncovered by our pipeline against other *B. bronchiseptica* strain genomes along with four genomes of closely related *Bordetella* species. We uncovered a total of 223 genes that are present in all *B. bronchiseptica* strains and absent in all other *Bordetella* species. To narrow down this set of potential PCR assay targets, we searched for the most conserved *B. bronchiseptica*-specific genes. As a point of reference, we determined the *sodC* gene used in the *N. meningitidis*-specific PCR assay (Kroll *et al.*, 1998) to be 99.6% identical among all six completely sequenced strains of *N. meningitidis*. There are seven *B. bronchiseptica*-specific genes with ≥99.6% sequence identity; these genes represent a prioritized list of potential PCR assay targets.

COMPUTATIONAL GENOMICS PIPELINE

We have presented our computational genomics pipeline, a local solution for automated, high-throughput computational support of prokaryotic genome sequencing projects. While the revolution in sequencing technology makes possible the execution of genome projects within individual laboratories, the computational infrastructure to fully realize this possibility does not yet exist. We made a comprehensive effort to put the tools required for this infrastructure into the hands of biologists working with next-generation sequencing data. Our aim in the course of this project was to facilitate decentralized biological discoveries based on affordable whole-genome prokaryotic sequencing, a mode of science termed 'investigator-initiated genomics'. For example, one project enabled by the pipeline in our laboratory is a platform for SNP detection and analysis in groups of bacterial genomes.

One of our major goals was to provide full automation of our pipeline's entire workflow, and this has been achieved. On the other hand, to allow computationally savvy users to realize the power of customizability, a semi-automated process is desirable. We have made an effort to

strike a balance between these objectives, and provide a modular, hierarchically organized structure to permit maximum customization when so desired.

The state of the art in prokaryotic computational genomics moves at a formidable pace. The modular organization of our pipeline, along with the emphasis on integration of complementary software tools, allows us to continually update our platform to keep pace with developments in computational genomics. For instance if a new, better assembler becomes available, we can include its results in the assembly stage with a simple change to the pipeline code.

CG-Pipeline has been in made available to the public since June 2010. Since then, several institutions have used it to analyze countless genomes. Those institutions that use CG-Pipeline and are known to us are: Jordan Lab at Georgia Institute of Technology and MVPDB (ourselves), Pacific Biosciences, National Biodefense Analysis and Countermeasures Center under the Department of Homeland Security, The Hong Kong Bioinformatics Centre, the core facilities at both CDC and Emory University, the University College Dublin, Scottish Crop Research Institute, and the Spain and Konstantinidis labs at the Georgia Institute of Technology. Internally in our collaboration between the CDC Meningitis and Vaccine Preventable Diseases Branch and The Jordan Lab, we have analyzed 23 *N. meningitidis* and *B. bronchiseptica* genomes. Thus the distribution and usage of CG-Pipeline is growing rapidly.

FUTURE UPDATES

CG-Pipeline is continuously being developed. As CG-Pipeline is modular, so are the updates for it. Installation for CG-Pipeline is automated. It places its executable scripts into an appropriate directory, downloads appropriate databases, and formats the databases as appropriate. However, we have noted that on some machines, installation might crash due to local customizations or different versions of prerequisite software. To address these cases, we will introduce a script that will check for all prerequisite software. If a user does not have a

particular software package, the script will provide brief instructions on how to obtain it. Secondly we will modify the installation script such that it will resume where it stopped in the case of a crash. One major reason for these crashes has been a reinvention of BLAST called BLAST+ which is quickly becoming standard and has different executables (Camacho *et al.*, 2009). CG-Pipeline will accommodate BLAST+ to avoid such crashes in the future.

Some users of CG-Pipeline have requested that Illumina data be valid input in addition to 454 SFF files. In response we have begun development for the incorporation of the Velvet (Zerbino & Birney, 2008) and Bowtie (Langmead *et al.*, 2009) short read assemblers. The assemblies from these programs will be incorporated into a combining stage as earlier described.

Although Minimus2 is a more optimal solution than most other combining tools, the combining stage can be optimized. Using an assembler like Minimus2 results in an inherent flaw. In the combination of two assemblies, a given read might be overrepresented and might bias an assembly output, or, a single read might be represented in two different locations. We have noted a tool specifically for the combination of two assemblies called the Reconciliator (Zimin *et al.*, 2008). In short, this tool detects disagreements between two assemblies (*e.g.* from AMOS and Newbler) and attempts a reassembly at these confounding sites. This represents the best method of combining two assemblies and is a necessity for the future of CG-Pipeline.

Although the feature prediction stage currently predicts with more than 95% accuracy, we have some improvements that can be made. It was noted that RNAmmer predicts start and stop sites within 0 to 9 nt of the correct sites. Using BLAST to fine-tune these start and stop sites would increase the overall accuracy of rRNA prediction. Furthermore BLAST would give higher confidence to all rRNA predictions. These BLAST databases will be obtained from an

rRNA curation site such as the Ribosomal Database Project (Cole *et al.*, 2009). Secondly for the prediction stage, coding sequence prediction accuracy can be increased if a Gibbs sampler were used. Gibbs sampling is a useful step for GeneMark (Lukashin & Borodovsky, 1998), which will help in finding start sites in coding sequences by training GeneMark's hidden markov model.

Feature Annotation can be made better by introducing subcellular localization. To that end, many packages offer some predictions (Bendtsen *et al.*, 2004; Krogh *et al.*, 2001; Yu *et al.*, 2010). However, one group has already created a decision-making tree using these tools and more for subcellular prediction (Emanuelsson *et al.*, 2007). We plan to incorporate this algorithm such that CG-Pipeline will be capable of subcellular localization. Subcellular localization is an important annotation. One example is the prediction of all surface antigens on a pathogen's surface, such that vaccine targets can be predicted—a process called reverse vaccinology (Rappuoli, 2000). One last improvement will be to incorporate genome-wide statistics. For many genome projects, it is an important task to categorize the functions of all genes, usually by Gene Ontology (GO) or Clusters of Orthologous Genes (COGs) (Ashburner *et al.*, 2000; Tatusov *et al.*, 2003). Therefore we will create an inventory of ratios of genes for each category in GO or COGs at the end of the annotation stage. Additionally, genome statistics at each of the three stages will be outputted with information such as the total number of coding genes, noncoding genes, and assembly statistics (*e.g.* number of contigs).

ACKNOWLEDGEMENTS

We are grateful to all participants of the Georgia Tech Computational Genomics class; to Leonardo Mariño-Ramírez for valuable guidance and input; and to Joshua S. Weitz for his support.

FUNDING

This work was supported by Defense Advanced Research Projects Agency [A.O.K. by HR0011-05-1-0057]; The Alfred P. Sloan Foundation [I.K.J. by BR-4839]; Georgia Research Alliance [I.K.J., P.J., S.A. by GRA.VAC09.O]; Centers for Disease Control and Prevention [L.S.K. by 1 R36 GD 000075-1]; and Bioinformatics program, Georgia Institute of Technology [J.H., P.J, V.N., S.A.].

CHAPTER 5

NEISSERIA BASE: A COMPARATIVE GENOMICS DATABASE FOR *NEISSERIA MENINGITIDIS*

ABSTRACT

Neisseria meningitidis is an important pathogen, causing sudden and life-threatening diseases including meningitis, septicemia and in some cases pneumonia. Genomic studies hold great promise for the future of *N. meningitidis* research, but substantial database resources are needed to deal with the wealth of information that comes with completely sequenced and annotated genomes. To address this need, we introduce *Neisseria* Base (NBase), which is a comparative genomics database and genome browser that houses and displays all publically available *N. meningitidis* genomes. In addition to existing *N. meningitidis* genome sequences, we sequenced and annotated 19 new genomes using 454 pyrosequencing and the CG-Pipeline genome analysis tool. In total, NBase hosts 27 complete *N. meningitidis* genome sequences along with their associated annotations. The NBase platform is designed to be scalable, via the underlying database schema and modular code architecture, such that it can readily incorporate new genomes and their associated annotations. The front page of NBase provides user access to these genomes through searching, browsing and downloading. NBase search utility includes BLAST based sequence similarity searches along with a variety of semantic search options. All genomes can be browsed using a modified version of the GBrowse platform, and a plethora of information on each gene can be viewed using a customized details page. NBase also has a whole-genome comparison tool that yields single nucleotide polymorphism (SNP) differences between user-defined groups of genomes. To demonstrate how the SNP comparison tool can

be used to address biological questions, we have compared ST-11 genomes against other genomes to identify markers.

Database URL: http://nbase.biology.gatech.edu

INTRODUCTION

MENINGOCOCCAL DISEASE

Neisseria meningitidis is a gram-negative encapsulated bacterium that is a leading worldwide cause of bacterial meningitis (Rosenstein *et al.*, 2001). Meningococcal meningitis and sepsis can cause death within hours and are particularly lethal in young children and adolescents. Meningitis case fatality rates range from 10 to 14%, and many survivors have long term neurological sequelae such as deafness and mental retardation. Each year, there are two to three thousand cases of meningococcal meningitis in the United States with about a 10% case fatality rate. Understanding the genomics of circulating strains is important for understanding the population biology of *N. meningitidis*. For example, a recent database BIGSdb takes advantage of loci across the meningococcal genome to produce customizable molecular profiles which reveals high-quality resolution typing data (Maiden & Jolley, 2010). Our needs include searching whole genomes for genomic determinants for phenotypic differences between isolates. To meet our needs, the Meningitis Laboratory at the Centers for Disease Control and Prevention has adopted a genomics approach to the study of *N. meningitidis*.

GENOMICS AND BIOINFORMATICS FOR *N. MENINGITIDIS*

There are a number of efforts underway to characterize *N. meningitidis* genome sequences, and the amount of genomic data for this organism will increase exponentially in the near future. CDC recently used next-generation sequencing technology (Margulies *et al.*, 2005) to characterize 19 *N. meningitidis* genomes, and at the time of this writing there are 9 additional *N. meningitidis* genome sequences that have been characterized elsewhere (Joseph *et al.*, 2010;

Kislyuk *et al.*, 2010; Parkhill *et al.*, 2000; Peng *et al.*, 2008; Rusniok *et al.*, 2009; Schoen *et al.*, 2008; Tettelin *et al.*, 2000).

It is essential to develop bioinformatics tools that can handle these data and have the capacity to scale sufficiently to accommodate the coming flood of *N. meningitidis* genome sequences. It is also important to provide computational genomics applications that are accessible and useful to working biologists. Our group has previously addressed one aspect of these challenges via the development of a fully automated analytical pipeline that takes genome sequence data and sequentially performs genome assembly, gene prediction and functional annotation – the CG-pipeline (Kislyuk *et al.*, 2010). This allows investigators to gain rapid access to annotations for individual *N. meningitidis* genomes without laborious manual analysis. However, once such data are in hand, researchers will still need a way to visualize the information and to compare annotation data and sequences among different genomes. For example, it could prove informative to compare genome sequences between strains of *N. meningitidis* with distinct disease causing capacities or between isolates from different outbreaks. It will also be critical to develop and maintain a shared platform for the storage and dissemination of the data and results generated by the *N. meningitidis* genome projects. To address these aims, we have developed *Neisseria* Base (NBase), an online platform for the storage, dissemination and comparative analysis of *N. meningitidis* genomes characterized at CDC and elsewhere. NBase allows users to browse, search and download genome sequences and annotations for *N. meningitidis*. The database also includes comparative genome analysis applications including the 'SNPtool' that allows users to discover individual nucleotide variations that distinguish between user-selected groups of *N. meningitidis* genomes. We show an example whole genome comparison in this manuscript. The NBase platform is designed to be scalable such that it can readily incorporate scores of new genomes and their associated

annotations. NBase is a freely available community resource that can be found at http://nbase.biology.gatech.edu.

MATERIALS AND METHODS

GENOMIC DATA

A total of 19 *N. meningitidis* genomic sequences were characterized via 454 pyrosequencing at CDC Biotechnology Core Facility Branch and analyzed at Georgia Tech. An additional 8 previously characterized *N. meningitidis* genomes are also included in NBase. A list of the genomes in NBase, along with metadata describing their origins, can be found in Table C.1. Genomic sequence data characterized at CDC were analyzed using the CG-Pipeline (versions 0.2.1 to 0.2.4) automated genome analysis platform. The CG-Pipeline was previously developed by our group for the automated assembly and annotation of prokaryotic genome sequences (Kislyuk *et al.*, 2010), and it is freely available at http://nbase.biology.gatech.edu. During the assembly stage the best reference genome assembly or best *de novo* assembly was chosen for further analysis. On the resulting assembly, gene locations were predicted using a combination of homology searches and *ab initio* methods. On those gene predictions, we performed automated functional annotation using a combined approach that includes 17 different annotation applications. The annotation step produces GenBank format flatfiles, and these GenBank files are converted into general feature format (GFF) for import into NBase. Further details of the genome sequencing protocol and the genome analysis procedures can be found in the report on the CG-Pipeline (Kislyuk *et al.*, 2010).

SOFTWARE COMPONENTS

To construct NBase, we used several software components: GBrowse 2.00 (Stein *et al.*, 2002), BLAST version 2.2.17 (Altschul *et al.*, 1997), Perl 5.8.8, BioPerl 1.6.1. (Stajich *et al.*, 2002), Mauve 2.2.0 (Darling *et al.*, 2004), MUSCLE 3.6 (Edgar, 2004) and JalView 2.5.1 (Clamp *et al.*,

2004; Waterhouse *et al.*, 2009). NBase rests on MySQL version 5.0 and is hosted using the Apache version 2.0 web server application (http://mysql.com, http://apache.org).

MULTIPLE SEQUENCE ALIGNMENT

Whole genome multiple sequence alignments (MSAs) were constructed from 27 genomes (Table C.1) using the program MAUVE with default settings. MAUVE produces local alignments of co-linear orthologous regions, shared among all genomes, which are called local co-linear blocks (LCBs). Each individual LCB was aligned further using the program MUSCLE, without using the refine option.

DETERMINATION OF SEQUENCE TYPES

Sequence types (STs) for some new genomes were determined by conventional sequencing methods (Maiden *et al.*, 1998). However for most new genomes because 454 pyrosequencing had already been performed, STs were determined using the whole genome sequence. Whole genomes were compared against the PubMLST database (Jolley *et al.*, 2004) to find and perform allele calls for all seven loci. The only ambiguous sequence was *adk* from M16917, which was resequenced to confirm its identity. Otherwise there were no novel alleles and no imperfect matches for these genomes.

COMPARISON OF ST-11 GENOMES AGAINST OTHER GENOMES

ST-11 genomes FAM18, M13519, M16917, M17661, M18774, M15141, and M9261 were placed into group 1. All other genomes in NBase were placed into group 2: Z2491, M13220, α14, α153, α275, M11791, M17094, M10699, M15293, M5178, MC58, M16207, 8013, M17062, M20899, M14900, M18575, 053442, M17277, and M20918 (Table C.1). MC58 was used as a reference genome. These groups of genomes were used in SNPtool (Figure 5.2), which in turn yielded SNPs that discriminate genomes between the two groups. These discriminating SNPs were given as coordinates on the MC58 genome.

NEISSERIA BASE

FRONT PAGE

The NBase front page serves as the gateway to all of the data, tools and analytical capability housed in NBase. The front page of NBase is designed to be straightforward. The user may choose to do one of several things from the front page: 1) view metadata, 2) download, 3) browse, 4) search, or 5) perform single nucleotide polymorphism (SNP) analysis.

Metadata, the information about each genome, is located on the left navigation bar (Figure 5.1C). Information about each isolate, such as geographic origin, date isolated, and profile information is shown in columns. Genomic data is available here to download in standard file formats (GenBank, FASTA, GFF). In addition, all software generated by this comprehensive project including the CG-Pipeline can be downloaded from the home page. On the Alignment Viewer page, the multiple sequence alignment of all available genomes may be viewed per LCB by either reading the plain text in Clustal format or using the Jalview applet (Clamp *et al.*, 2004; Waterhouse *et al.*, 2009). Jalview includes functions for generating phylogenetic trees on demand, which is useful for observing the similarity between strains in a specific region.

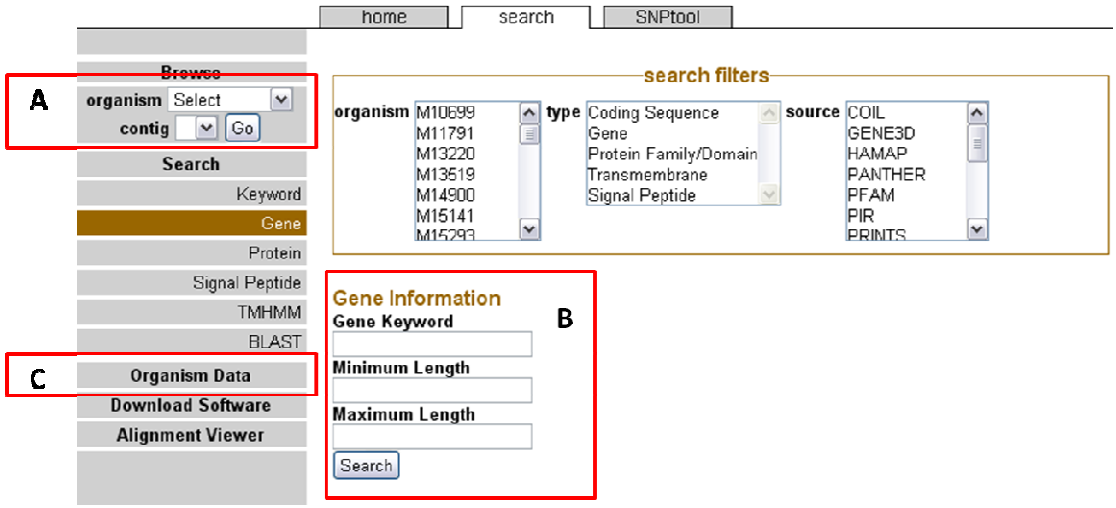


Figure 5.1. The Front Page Sidebar. The sidebar provides access to the genome browser, search functions, and metadata. (A) To browse, users can select a genome and a contig to proceed to the GBrowse interface. (B) To search a user can supply a keyword or can choose to search via one of several search items. (C) Genomic metadata is available via the Organism Data link (See Table C.1). A link is provided to the alignment data page.

Users can browse genomes by selecting the organism from the drop-down menu on the sidebar (Figure 5.1A). After a genome has been selected, available contigs or chromosomes are shown. Most 454 assemblies will not be complete due to coverage considerations (Lander & Waterman, 1988) or due to repeat elements (Pop *et al.*, 2004) and therefore most of the genomes on NBase are viewable on the contig level instead of a chromosomal level. After the contig or chromosome is chosen, the user is brought to the GBrowse (Stein *et al.*, 2002) graphical interface (see Genome Browser section).

Another way to arrive at the GBrowse interface is to search for specific genomic features. In this case, genomic features are any annotated landmarks in a genome including genes and all associated annotations. Categories of searches are located on the sidebar on the Search page (Figure 5.1B). Search results may be filtered to selected organisms, feature types (*e.g.* genes), and/or annotation source. The annotation sources correspond to different annotation

applications used in the CG-Pipeline genome analysis tool (Kislyuk *et al.*, 2010). Each specific type of search allows for the inclusion of more specific parameters such as gene length. If a user does not want to use a specific search, a general keyword search is provided. One more way to search for genomic sequences is to use any one of the five BLAST programs in the BLAST interface. By clicking on the BLAST results, the user can arrive at the GBrowse interface.

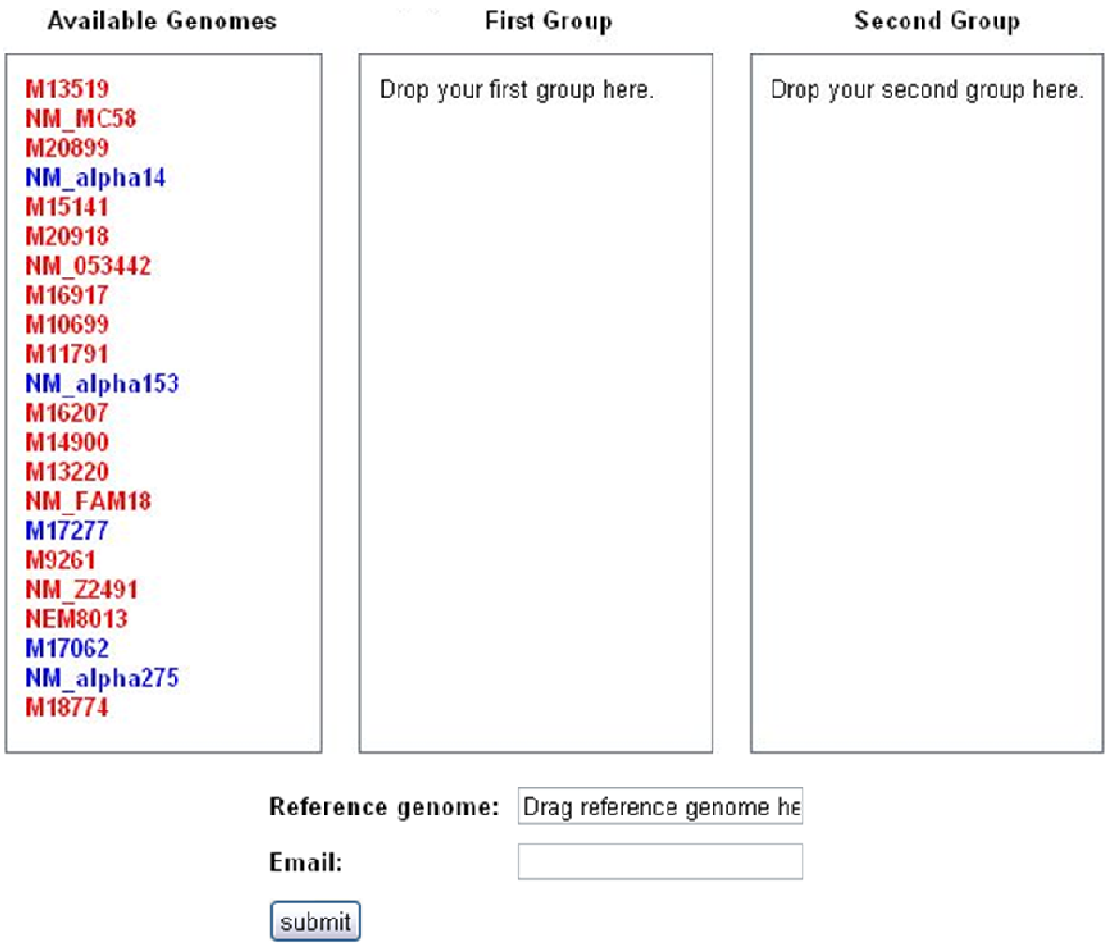


Figure 5.2. SNPtool. SNPtool finds discriminating SNPs between two groups of genomes. Each group is defined by the user, by dragging each genome to a designated group. Not all genomes must be used. A reference genome must be designated, as the results can be viewed on the graphical genome browser from the vantage point of the chosen reference genome. Invasive isolates are designated by red, carried by blue.

The user may navigate to the SNP analysis tool, called SNPtool, from the SNPtool tab at the top of the front page. SNPtool uses a comprehensive multiple sequence alignment (MSA) to discover SNPs that show mutually exclusive patterns between two groups of genomes (Figure 5.2). Such SNPs serve as markers for discriminating between the two groups of genomes. To use SNPtool, a user drags desired genomes into the first or second group to define each group.

Next, the user supplies a reference genome so that the results can be visualized in the GBrowse interface. The SNPtool outputs the coordinates of all discriminating SNPs, defined in the genome space of the reference genome, along with a list of genes associated with the discriminating SNPs. These SNP genes are defined as genes that have discriminating SNPs within their coding regions or within +/- 1kb of the predicted coding start/stop sites. All results from previous SNPtool selections are accessible by a hyperlink on the SNPtool page, and the user can view selected results in the Genome Browser.

GENOME BROWSER

We chose to use GBrowse because it is customizable, open-source, fast, and reliable (Stein *et al.*, 2002). Also it is a proven genome browser used by several institutions for several other organisms and so many users are familiar with the interface (Bieri *et al.*, 2007; Drysdale, 2008; Elsik *et al.*, 2006) (GMOD Users, http://gmod.org/wiki/GMOD_Users). GBrowse works with a MySQL database, using the GFF schema. We were able to use this schema unmodified to store multiple genome annotations in a single database and quickly load the feature annotations with GBrowse. Use of this approach also facilitates scalability with respect to rapid and facile assimilation of new genome sequences and annotations.

NBase is also designed to be scalable with respect to the addition of new applications. This has been achieved via the formatting of the database schema that underlies GBrowse. Currently, all of our search utilities, both text and sequence based, query the same MySQL database on which GBrowse runs. When any new applications are added to the site, they will be designed to query the same database with little or no modification to the schema. Furthermore, all of the source for the browser and associated utilities has a modular design to facilitate future additions to the site.

GBrowse gives the user a linear map of a selected region of a genome with the genomic features appearing at their respective coordinates (Figure 5.3). The user can zoom in and out, move upstream and downstream along the sequence, and configure the display of sequence features. GBrowse uses semantic viewing which refers to how much detail is shown when the depiction is zoomed in or zoomed out. From a distance, features are shown only as colored arrows which show directionality. At a medium zoom, features' names are visible. At a very close distance, individual nucleotides and amino acid residues can be seen.

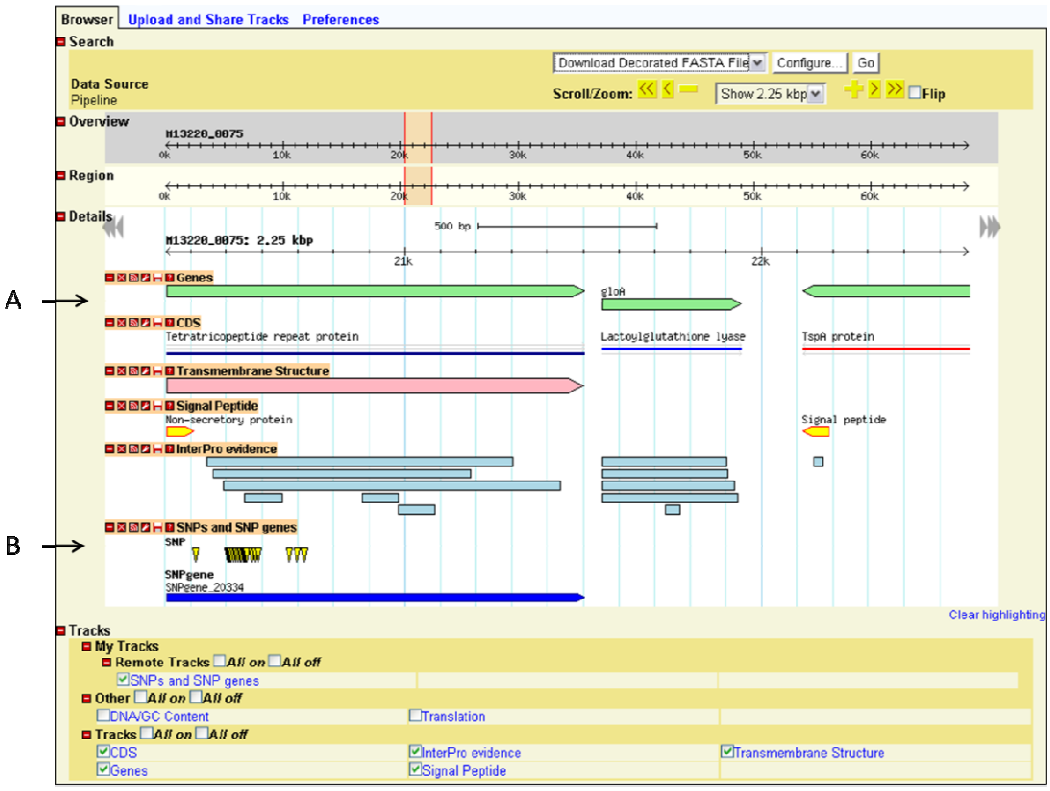


Figure 5.3. Genome Depiction. The genome is represented linearly, with features on their respective coordinates. (A) Genes and their CDS each have their own track and are links to their own details page. (B) Uploaded tracks.

Genes, nucleotides, and residues can only be viewed if their respective tracks have been turned on. Additionally users can supply their custom tracks to overlay and compare features, using the correct file format (Figure C.1). One example of adding custom tracks is using the results from SNPtool. The results of SNPtool, the lists of discriminating SNPs and SNP genes, can be uploaded to NBase and compared to other tracks (Figure 5.3).

DETAILS PAGE

Clicking on a feature brings the user to a details page (Figure 5.4). Each feature inherently has some fundamental characteristics that will be displayed: name, length, and coordinates. For

data that is produced by the CG-Pipeline or that is already present in imported genomes, there is a wealth of additional information available. The CG-Pipeline annotates genes using the UniProt database (2010) and InterProScan (Mulder & Apweiler, 2007). In addition, it predicts transmembrane helices with TMHMM (Krogh *et al.*, 2001), signal peptides with SignalP (Bendtsen *et al.*, 2004), and virulence genes with the virulence factors database (Chen *et al.*, 2005; Yang *et al.*, 2008). All of these annotations are present in NBase. The details page includes hyperlinks to the online UniProt and InterProScan databases from which the annotation is derived.

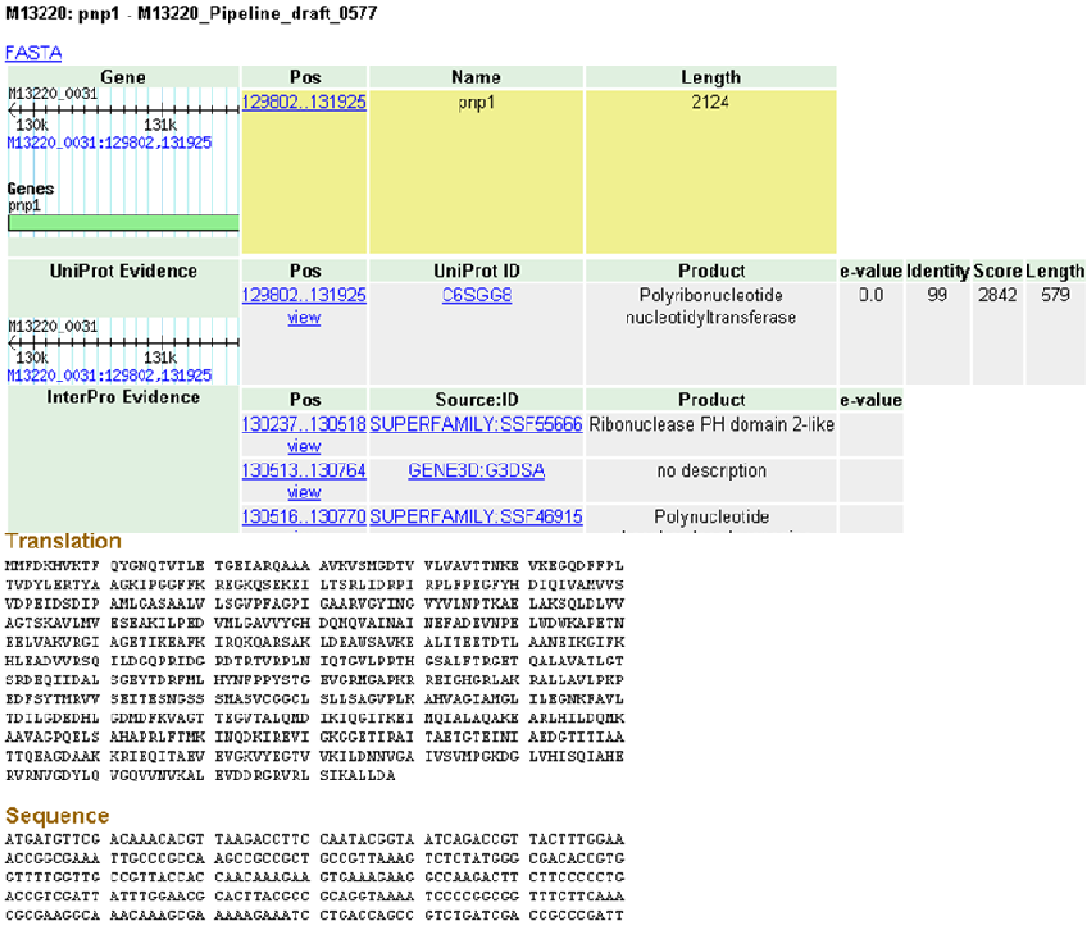


Figure 5.4. Details Page. The details for some features of a coding sequence are shown. Up to five overlapping feature types may appear: gene, protein, protein domain, signal peptide, and transmembrane structure. The target feature is highlighted in yellow. The nucleotide and amino acid sequences for this feature appear at the bottom of the page. All features on the page include links to their coordinates in GBrowse genome viewer.

BIOLOGICAL DISCOVERY USING SNPTOOL

Here, we illustrate the potential application of NBase, and the comparative genomics utility encoded therein, to biological discovery in *N. meningitidis*.

SEQUENCE TYPE 11

To molecularly profile and type meningococci, multilocus sequence typing (MLST) is used (Maiden *et al.*, 1998). In MLST, seven predefined loci are sequenced and their alleles are called.

Those allele calls are concatenated to produce a profile called a sequence type (ST). At the time of this writing, several hundred alleles are defined per locus, and more than 8,500 STs are defined (Jolley *et al.*, 2004).

Since its inception, MLST has been used to define the population structure of *N. meningitidis*. Consequently, some STs have been correlated with phenotype, especially those correlated with disease. One influential study found that ST-11 meningococci are more associated with disease than any other ST (Yazdankhah *et al.*, 2004). We compared all available ST-11 meningococcal genomes against all other available genomes (Table C.1). The aim of this analysis is to show markers for ST-11 outside of the seven predefined loci of MLST (Maiden *et al.*, 1998). Whole genome comparison yielded 2,589 SNPs that show mutually exclusive patterns between ST-11 genomes and all others available to us. All of these markers represent possible markers for ST-11, which by definition could detect ST-11 apart from other STs. While MLST is used on a kilobase magnitude, this SNP analysis represents a resolution at 3 fold finer magnitude and could be used to find smaller regions or even nucleotides that distinguish an ST from another. This is a proof-of-principle that shows that the SNPtool is able to identify differences between groups of genomes at a very fine nucleotide resolution. Theoretically the SNPtool could also be used to match phenotypes to genotypes. For instance, if the genomes of all fast-growing meningococci were compared against slow-growing meningococci then markers for its growth, and even putatively causative genes, could be identified.

DISCUSSION

Although there are other neisserial databases, *e.g.* (Aurrecoechea *et al.*, 2007; Dehal *et al.*, 2010; Flicek *et al.*, 2010; Kent *et al.*, 2002; Sayers *et al.*, 2010), NBase is customized for the CG-Pipeline (Kislyuk *et al.*, 2010). Other genome browsers have great strengths, especially NeMeSys (Rusniok *et al.*, 2009) which provides very useful tools, provides a syntenic perspective

when viewing genomes, and contains mutagenesis studies for genome 8013 (Geoffroy *et al.*, 2003). However, NBase is the genome browser that can easily accommodate new genomes that are analyzed using the CG-Pipeline, which is open source and can run locally on a desktop computer.

Using only open source software has many advantages. First, the GBrowse interface is familiar to many, and so the learning curve is lessened. Second, all new plugins for GBrowse are readily incorporated into NBase as needed (*e.g.* a plugin to download a selected region of a genome). Third, all software belonging to the Generic Model Organism Database project (GMOD) can be assimilated into NBase. For example in the future, we will be incorporating GMOD's SynView which is a synteny viewer for GBrowse (Wang *et al.*, 2006).

NBase allows searching, browsing, and downloading of whole genomes (including assemblies and annotations). In total we have 27 meningococcal genomes available for these tasks. At NBase's core is GBrowse which has a simple, yet sophisticated, database structure and which has a graphical interface for browsing genomes. GBrowse also facilitates the usage of a details page for each feature in a genome so that, not only does a user see the breadth of a genome, but the user can also see depth. Finally, we have demonstrated the malleability and scalability of NBase by incorporating a custom tool, SNPtool. SNPtool compares two groups of genomes and displays discriminating SNPs and their associated SNP genes. Furthermore, we have demonstrated that SNPtool can be used to compare whole groups of isolates on a genomic level to uncover individually significant SNPs. This has been useful for uncovering genomic markers for ST-11.

FUNDING

This work was supported by Centers for Disease Control and Prevention [1 R36 GD000075-1 to L.S.K.]; Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular

Biology [BR-4839 to I.K.J.]; Georgia Research Alliance [GRA.VAC09.O to I.K.J., B.H.H., L.W.M., and L.S.K.]; Bioinformatics program, Georgia Institute of Technology [to J.C.H., P.J., N.V.S., V.N.]; and Defense Advanced Research Projects Agency [HR0011-05-1-0057 to A.O.K.].

ACKNOWLEDGEMENTS

Thanks go to Elizabeth Neuhaus, Dhwani Govil, and Scott Sammons from the Biotechnology Core Facility Branch who helped guide us. Thank you to Scott Cain who originally demonstrated the GBrowse platform to us. Thank you to the 2008, '09, and '10 Compgenomics classes at The Georgia Institute of Technology. Thank you to Nancy Messonnier for reviewing this manuscript and providing valuable feedback. Thank you to the many individuals who donated strains for research.

CHAPTER 6

USING SNPS TO DISCRIMINATE DISEASE ASSOCIATED FROM CARRIED GENOMES OF *NEISSERIA MENINGITIDIS*

ABSTRACT

Neisseria meningitidis is one of the main agents of bacterial meningitis causing substantial morbidity and mortality worldwide. However, most of the time *N. meningitidis* is carried as a commensal not associated with invasive disease. The genomic basis of the difference between disease associated versus carried isolates of *N. meningitidis* may provide critical insight into mechanisms of virulence, yet it has remained elusive. Here, we have taken a comparative genomics approach to interrogate the difference between disease associated and carried isolates of *N. meningitidis* at the level of individual nucleotide variations (*i.e.*, SNPs). We aligned complete genome sequences of 8 disease associated and 3 carried isolates of *N. meningitidis* to search for SNPs that show mutually exclusive patterns of variation between the two groups. We found 801 SNPs that distinguish the 8 disease associated from the 3 carried genomes of *N. meningitidis*, which is far more than can be expected by chance alone given the level of nucleotide variation among the genomes. The putative list of disease associated versus carriage discriminating SNPs may be expected to change with increased sampling or changes in the identities of the isolates being compared. Nevertheless, we show that these discriminating SNPs are more likely to reflect phenotypic differences than shared evolutionary history. Discriminating SNPs were mapped to genes and the functions of the genes were evaluated for possible connections to virulence mechanisms. A number of over-represented functional categories related to virulence were uncovered among SNP genes including: oxidoreductases, cell wall, immune system evasion and proteases.

INTRODUCTION

Neisseria meningitidis, the meningococcus, is a leading cause of bacterial meningitis worldwide with devastating morbidity and mortality (Centers for Disease Control and Prevention [http://www.cdc.gov/meningitis/about/faq.html]). *N. meningitidis* is a gram-negative encapsulated diplococcus of the human nasopharynx that is most frequently found to be asymptomatically carried (Maiden & Caugant, 2006); ~10% of healthy individuals are carriers of *N. meningitidis* (Claus *et al.*, 2005; Dolan-Livengood *et al.*, 2003). In a small number of carriers, some strains of *N. meningitidis* are able to invade epithelium of the nasopharynx and enter the bloodstream, thus leading to invasive disease, such as meningococcal meningitis and meningococcemia (Rosenstein *et al.*, 2001).

A number of previous studies have taken a comparative genomics approach to try and determine if there is any genomic basis for the difference between disease associated and carried isolates of *N. meningitidis*. Perrin *et al.* (2002) used comparative genome hybridization (CGH) to compare the genomes of disease associated *N. meningitidis* strains against the genomes of the closely related species *N. gonorrhoeae* and *N. lactamica* (Perrin *et al.*, 2002). They were able to find a number of chromosomal regions present only in *N. meningitidis*, suggesting a possible role in species-specific virulence. However, genes found in the species-specific regions would later be shown to be present in both disease associated and carried isolates of *N. meningitidis* (Schoen *et al.*, 2008). A subsequent CGH study discovered 55 genes present in all *N. meningitidis* serogroup B isolates analyzed and absent in *Neisseria* commensal species (Stabler *et al.*, 2005). Nevertheless, several of these serogroup B strains were carried isolates that were not associated with disease. It was also shown later that the majority of genes previously implicated in virulence using the comparative approach were shared between *N. meningitidis* and the non-pathogenic *N. lactamica* (Snyder & Saunders, 2006).

In 2005, Bille *et al.* discovered an 8kb bacteriophage-derived sequence shared among 29 disease associated *N. meningitidis* genomes and largely absent from carried isolates (Bille *et al.*, 2005). While no single gene in this prophage met the condition of being present in all disease associated genomes and absent in all carried genomes, the distribution of prophage genes was highly skewed towards disease associated genomes. Thus, at that time, this genetic island represented the best example of a genomic feature that could distinguish disease associated from carried genomes of *N. meningitidis*. However, the next year Hotopp *et al.* used a more exhaustive CGH study to show that this prophage was actually found in the genomes of 60% of disease associated isolates and 42% of carried isolates (Hotopp *et al.*, 2006).

In 2008, Schoen *et al.* performed the first complete genome sequence based comparison between disease associated and carried isolates of *N. meningitidis* (Schoen *et al.*, 2008). These authors found that genes previously implicated in virulence were widely shared among disease associated and carried genomes; in other words, there does not appear to be any core pathogenome for *N. meningitidis*. Later, the same group used genome sequence comparison between a disease associated versus a carried isolate of serogroup B *N. meningitidis* strains, to show that virulence in *N. meningitidis* is likely to be encoded by sequence differences found across numerous genes (Joseph *et al.*, 2010). Taken together, the results of all of these comparative genomic studies indicate that the presence or absence of specific genes, sets of genes, or other large-scale genomic features cannot be used to distinguish disease associated from carried isolates of *N. meningitidis*.

In light of these previous results, we decided to explore the utility of individual nucleotide variations for discriminating between disease associated versus carried isolates of *N. meningitidis*. The use of nucleotide variation takes advantage of advances in sequencing technology to provide a deeper level of resolution for genome comparisons. We hypothesized

that single nucleotide polymorphisms (SNPs) will provide markers that can distinguish disease associated from carried isolates of *N. meningitidis*. To test this hypothesis, we compared complete genome sequences of 8 disease associated (Bentley *et al.*, 2007; Kislyuk *et al.*, 2010; Parkhill *et al.*, 2000; Peng *et al.*, 2008; Tettelin *et al.*, 2000) and 3 carried isolates (Schoen *et al.*, 2008) of *N. meningitidis*. The disease associated isolates were sampled from individuals with meningococcal disease, and the carried isolates were taken from asymptomatic individuals. Thus, the two groups of isolates represent instances of phenotypic differences in *N. meningitidis* virulence, and we sought to assess whether there may be genomic determinants of these differences. To do this, we searched for SNPs that show mutually exclusive patterns of variation between the two groups of isolates. We found that hundreds of SNPs can serve as markers that distinguish these sets of disease associated versus carried isolates of *N. meningitidis*, and these discriminating SNPs are more likely to reflect phenotypic differences than shared evolutionary history. We mapped these discriminating SNPs to *N. meningitidis* genes to assess their potential functional significance.

The transition from asymptomatic to disease associated states for *N. meningitidis* may be extremely rapid. Carried isolates studied here belong to a serogroup and ST combination that has been shown to cause a small percentage of meningococcal disease and isolates from disease associated strains spend most of their time being carried. Thus, the disease associated versus carried isolates studied here yield a snap-shot in time and place of a set of nucleotide variants that distinguish one group of *N. meningitidis* disease associated isolates from a group of carried isolates. Accordingly, the particular set of discriminating SNPs characterized here, along with the list of SNP genes, will likely change as additional genome sequences are characterized and compared.

MATERIALS AND METHODS

N. MENINGITIDIS CULTURE CONDITIONS AND DNA EXTRACTION

Isolates were stored at -80°C in defibrinated sheep blood (Lampire, Pipersville, Pennsylvania) prior to use, and were subsequently streaked onto Chocolate II Agar (BBL, Sparks, Maryland) and incubated at 37°C overnight with 5% CO_2 before harvesting for DNA preparation. Purified genomic DNA was extracted using the Blood and Cell Culture DNA Maxi Kit (Qiagen, Valencia, California) following the manufacturer's instructions. The DNA concentration and the 260/280 ratio was obtained using a NanoDrop ND-1000 Spectrophotometer (NanoDrop Products, Wilmington, Delaware).

GENOME SEQUENCING AND ANALYSIS

Sequencing of *N. meningitidis* isolates M9261, M13220, M10699, and M15141 was performed using Roche Applied Science/454 pyrosequencing in the CDC Biotechnology Core Facility; each strain was sequenced using the GS-20 platform. For each genomic DNA preparation, a random shotgun library was produced using Roche protocols for nebulization, end-polishing, adaptor ligation, nick repair and single-stranded library formation (Margulies *et al.*, 2005). Following emulsion PCR, DNA bound beads were isolated and sequenced using long read (LR) sequencing kits. Sequencing was followed by read trimming and refiltering to recover short quality reads.

Genome sequence assemblies were performed using a customized genome analysis pipeline (CG-pipeline version 0.2.1) (Kislyuk *et al.*, 2010) that combines reference based assembly using the Newbler assembler and AMOScmp (Pop *et al.*, 2004). Results from the two assemblers were combined using Minimus. The analytical pipeline was also used to perform gene prediction and functional annotation using a combination of tools. The four annotated genomes (Table 6.2)

were uploaded into a customized database, *Neisseria* Base (NBase), based on the GBrowse platform (Stein *et al.*, 2002).

GENOME ALIGNMENT AND SNP ANALYSIS

The four *N. meningitidis* isolates characterized here were analyzed together with 7 other completely sequenced *N. meningitidis* isolates (Table 6.1). Complete genome sequences were aligned using the program MAUVE, which locates and aligns conserved and syntenic genomic regions called local colinear blocks (LCBs) (Darling *et al.*, 2004). Default MAUVE alignment settings were used with the exception of a minimum LCB weight of 500. M13220 was set as the reference genome, whereby all other genomes were rearranged according to M13220. Only LCBs that contained conserved regions of all 11 isolates (*n*=69) were used for subsequent SNP analysis. The 69 individual LCBs were re-aligned using ClustalW (version 1.83) (Thompson *et al.*, 2002). LCB alignments were analyzed to look for discriminating nucleotide patterns (*i.e.* SNPs) separating the 8 disease associated genomes from the 3 carried genomes (Table 6.1). The disease associated isolates were sampled from individuals with meningococcal disease and the carried isolates were taken from asymptomatic individuals. Here, SNPs include nucleotide variations along with insertions and deletions (indels). A discriminating SNP is defined as a polymorphic site that shows one nucleotide pattern for one group of genomes (disease associated or carried), and a mutually exclusive pattern for the other group (Figure 6.1). The total number of disease associated versus carried SNPs was computed and compared to a null distribution of expected discriminating SNP counts, given the background polymorphism level, generated using simulation. For simulation, the identity of the nucleotides at each polymorphic site were randomly permuated among genomes. Polymorphic sites were then evaluated to come up with a count of discriminating SNPs for simulation. This process was repeated 10,000 times.

TABLE 6.1. *N. meningitidis* genomes compared in this study.

Name	Serogroup	ST[a]	CC[b]	Isolate Phenotype	Location (year)	Accession
M13220	A	ST-7	ST-5	Disease	Phillipines (2005)	SRS074220
M10699	B	ST-32	ST-32	Disease	Oregon (2003)	SRS074512
M15141	C	ST-11	ST-11	Disease	New York (2006)	SRS074514
M9261	W135	ST-11	ST-11	Disease	Burkina Faso (2002)	SRS074515
053442	C	ST-4821	ST-4821	Disease	China (2003)	CP000381
Z2491	A	ST-4	ST-4	Disease	Gambia (1983)	AL157959
FAM18	C	ST-11	ST-11	Disease	North Carolina (1980s)	AM421808
MC58	B	ST-74	ST-32	Disease	United Kingdom (1985)	AE002098
α14	*cnl*[c]	ST-53	ST-53	Carried	Bavaria (1999-2000)	AM889136
α153	29E	ST-60	ST-60	Carried[d]	Bavaria (1999-2000)	AM889137
α275	W135	ST-22	ST-22	Carried[d]	Bavaria (1999-2000)	AM889138

[a] Sequence Type

[b] Clonal Complex

[c] Capsule null locus (Claus *et al.*, 2002)

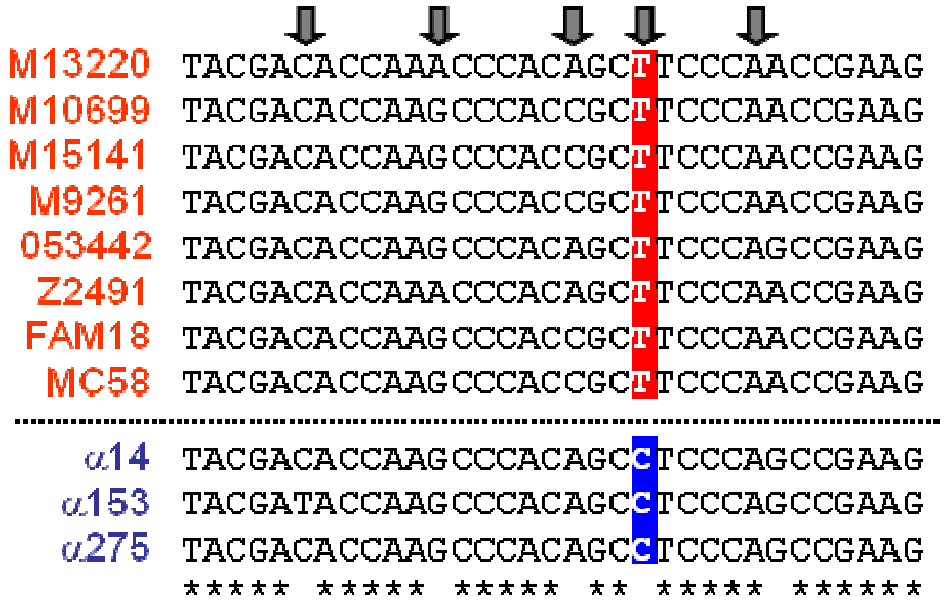
[d] The isolates studied here are taken from asymptomatic individuals (Claus *et al.*, 2005). Different isolates of this serogroup and ST combination have been shown to cause a small percentage of meningococcal disease in Bavaria (Schoen *et al.*, 2008).

TABLE 6.2. New *N. meningitidis* genomes recently characterized. These genomes are reported in their final form in (Kislyuk *et al.*, 2010).

Name	Location and date	Contigs	Assembly length [a]	Genes
M13220	Philippines Jan 2005	82	2.20	2108
M10699	Oregon May 2003	40	2.18	1978
M15141	New York City Aug 2006	50	2.28	2141
M9261	Burkina Faso Apr 2002	79	2.21	2261

[a] Given in megabases

A



B

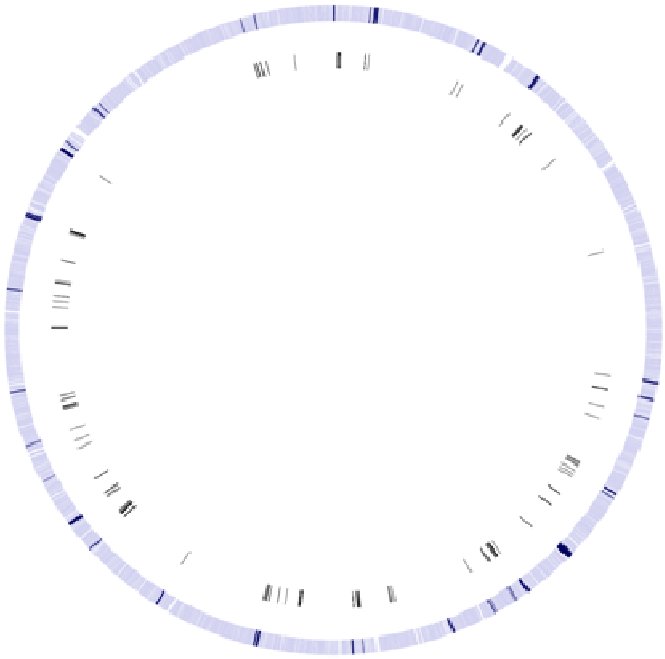


Figure 6.1. Disease associated versus carried isolate genome discriminating SNPs.

(A) Genome sequence alignment between disease associated (red) and carried (blue) isolates were analyzed for the presence of SNPs (gray arrows). An example of a disease associated versus carried genome discriminating SNP is highlighted. (B) A concatenated genome map of M13220. The outer ring shows all genes, with SNP genes labeled in black and others in blue. The inner ring shows the locations of SNPs. Continued on the next page.

C

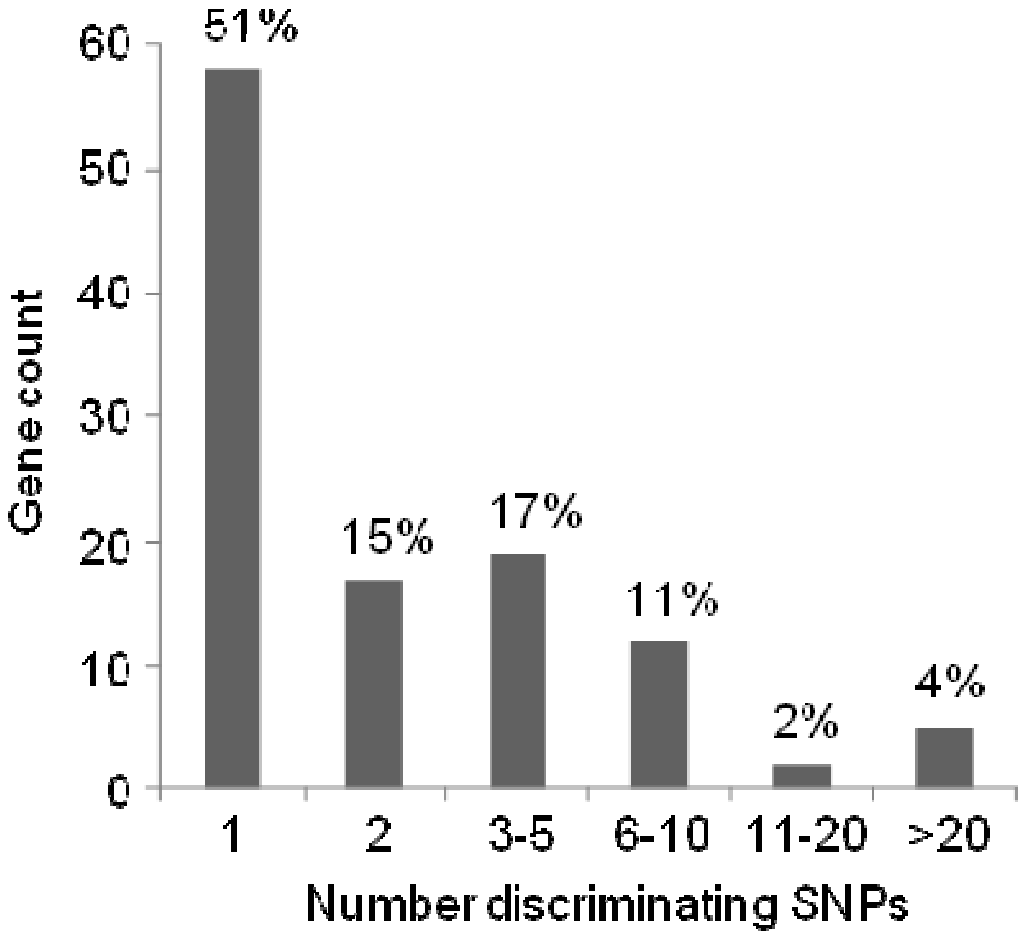


Figure 6.1 (continued). Disease associated versus carried isolate genome discriminating SNPs. (C) A histogram of the number of discriminating SNPs per SNP gene.

PHYLOGENETIC ANALYSIS

Whole genome sequence alignments were used to calculate nucleotide *p*-distances between *N. meningitidis* isolate genomes, and the distances were used to reconstruct an *N. meningitidis* phylogeny with the Neighbor-joining algorithm (Saitou & Nei, 1987) implemented in the program MEGA 4 (Kumar *et al.*, 2008). The same approach was used to reconstruct an *N. meningitidis* phylogeny based on a concatenated nucleotide sequence alignment of the 7 multilocus sequence typing (MLST) loci: *abcZ*, *adk*, *aroE*, *fumC*, *gdh*, *pdhC*, and *pgm* (Holmes *et al.*, 1999; Maiden *et al.*, 1998). 1,000 bootstrap replicates of the MLST alignment were used to evaluate the confidence of the phylogeny.

FIXATION INDEX (F_{ST}) ANALYSIS

The F_{ST} statistic (Hudson *et al.*, 1992) was used as a measure of the genetic differentiation between groups of *N. meningitidis* genomes. F_{ST} was measured using the pairwise nucleotide *p*-distances calculated from the *N. meningitidis* whole genome alignment. Disease associated and carried isolate groups were used to compute the average within group genome *p*-distance ($\prod within$) and the average between group genome *p*-distance ($\prod between$). F_{ST} was then calculated as $1-\frac{\prod within}{\prod between}$. Simulation was used to compute a background distribution of F_{ST} values that could be expected given the levels of nucleotide variation among all genomes. To do this, *N. meningitidis* genomes were randomly assigned to either the disease associated (*n*=8) or carried (*n*=3) groups and F_{ST} was re-calculated based on the random groups. This was repeated 10,000 times to yield a null frequency distribution of expected F_{ST} values.

BAYESIAN CLUSTERING METHOD

We used a Bayesian method implemented by the program STRUCTURE version 2.3.1 to determine the optimal number of groups (*K*) that best represent the underlying nucleotide

variation (*i.e.* the structure) among the *N. meningitidis* genomes analyzed here (Pritchard *et al.*, 2000). STRUCTURE was run with *K*=1, 2, 3, 4, and 5 groups. For each *K*, the burn-in and run length parameter values were set to 50,000 each.

GENOMIC AND FUNCTIONAL CHARACTERISTICS OF DISCRIMINATING SNPS

Discriminating SNPs were mapped to *N. meningitidis* gene locations – either internal to or within 1kb of the coding sequence. The resulting set of SNP genes (proteins) was then evaluated for statistically significant enrichment for a variety of functional characteristics including gene ontology (GO) annotations, the presence of a signal peptide, identity as a lipoprotein, horizontal transfer, subcellular location, and identity as a putative virulence factor. GO annotations were taken from the InterProScan database (Zdobnov & Apweiler, 2001). The presence of signal peptides were inferred using the SignalP program (Emanuelsson *et al.*, 2007). Lipoprotein status was inferred using the LipoP program (Juncker *et al.*, 2003). The horizontal transfer status of genes were inferred using a combination of three programs: BLAST (Altschul *et al.*, 1997), CodonO (Angellotti *et al.*, 2007) and AlienHunter (Vernikos & Parkhill, 2006), along with an analysis of GC-content. Genes were called as putative virulence factors based on the Virulence Factors Database (VFDB) (Chen *et al.*, 2005; Yang *et al.*, 2008). SNP genes were evaluated for statistical over-representation for each functional characteristic using the hypergeometric test implement in the program GeneMerge (Castillo-Davis & Hartl, 2003). The hypergeometric test in this study gives the probability *P* of selecting *r* genes with a functional characteristic in the set of SNP genes *k* from an overall set of genes in the genome *n*, where *p* is the proportion of *r* genes in the population and sampling is without replacement (Equation 6.1). SNP genes found to encode significantly over-represented functions were further evaluated using BLAST homology searches from three sources: our genome browser NBase

(http://nbase.biology.gatech.edu), NCBI's RefSeq database (Pruitt *et al.*, 2007), and the NeMeSys database (Rusniok *et al.*, 2009).

$$P(r \mid n, p, k) = \frac{C_{r}^{pn} C_{k-r}^{(1-p)n}}{C_{k}^{n}} \tag{6.1}$$

RESULTS AND DISCUSSION

COMPARATIVE GENOMIC SEQUENCE ANALYSIS OF *N. MENINGITIDIS*

We hypothesize that SNPs can be used as markers that distinguish sets of disease associated versus carried genomes of *N. meningitidis*. To test this hypothesis, and to search for potential genomic influences on virulence, we performed comparative sequence analysis of 11 isolates of *N. meningitidis* with completely sequenced genomes: 8 disease associated isolates from individuals with meningococcal disease and 3 carried isolates from asymptomatic individuals (Table 6.1). The 4 previously published disease associated genomes are taken from a series of individual genome projects and represent the most common disease associated *N. meningitidis* serogroups : A, B, and C (Bentley *et al.*, 2007; Parkhill *et al.*, 2000; Peng *et al.*, 2008; Tettelin *et al.*, 2000). The 3 previously published carried genomes were reported in 2008 as part of a comparative analysis of disease associated and asymptomatically carried *N. meningitidis* genomes that focused on differences in the presence and absence of virulence factor genes between the two groups (Schoen *et al.*, 2008). Although these three isolates were taken from asymptomatic individuals during a carriage study (Claus *et al.*, 2005), other isolates of this serogroup and ST combination have been shown to cause a small percentage of meningococcal disease in Bavaria (Schoen *et al.*, 2008). Recently, we reported the characterization of 4

additional disease associated genomes of *N. meningitidis* that cover serogroups A, B and C and also include the first reported disease associated W135 serogroup genome sequence (Kislyuk *et al.*, 2010). The W135 isolate characterized here was isolated in Burkina Faso and was the cause of a major outbreak of bacterial meningitis at the 2000 Hajj (Dull *et al.*, 2005; Lingappa *et al.*, 2003; Mayer *et al.*, 2002).

The *N. meningitidis* genomes were characterized via pyrosequencing using either a single or a half run on the Roche 454 instrument (Table 6.2). The number of reads produced in the 4 experiments ranged from 197,000-605,000, and the average read lengths were 105-245 base pairs. Altogether, these data yielded 47.6-94.3 million bases per genome amounting to 20-40x coverage for the ~2.2 megabase *N. meningitidis* genomes. We developed customized genome assembly, gene prediction and functional annotation pipelines to analyze these data (Kislyuk *et al.*, 2010). Our genome assembly procedure resulted in an order-of-magnitude decrease in the number of contigs produced by the Newbler assembler that ships with the 454 platform. There are 38-82 contigs that cover 2.18-2.28 megabases of the 4 genomes. We annotated 1978-2261 genes including protein coding genes, non-coding RNAs and insertion sequences. All 4 of the new genomes reported here, along with custom annotations and tools for searching and comparative sequence analysis, are available at our genome browser database (NBase http://nbase.biology.gatech.edu).

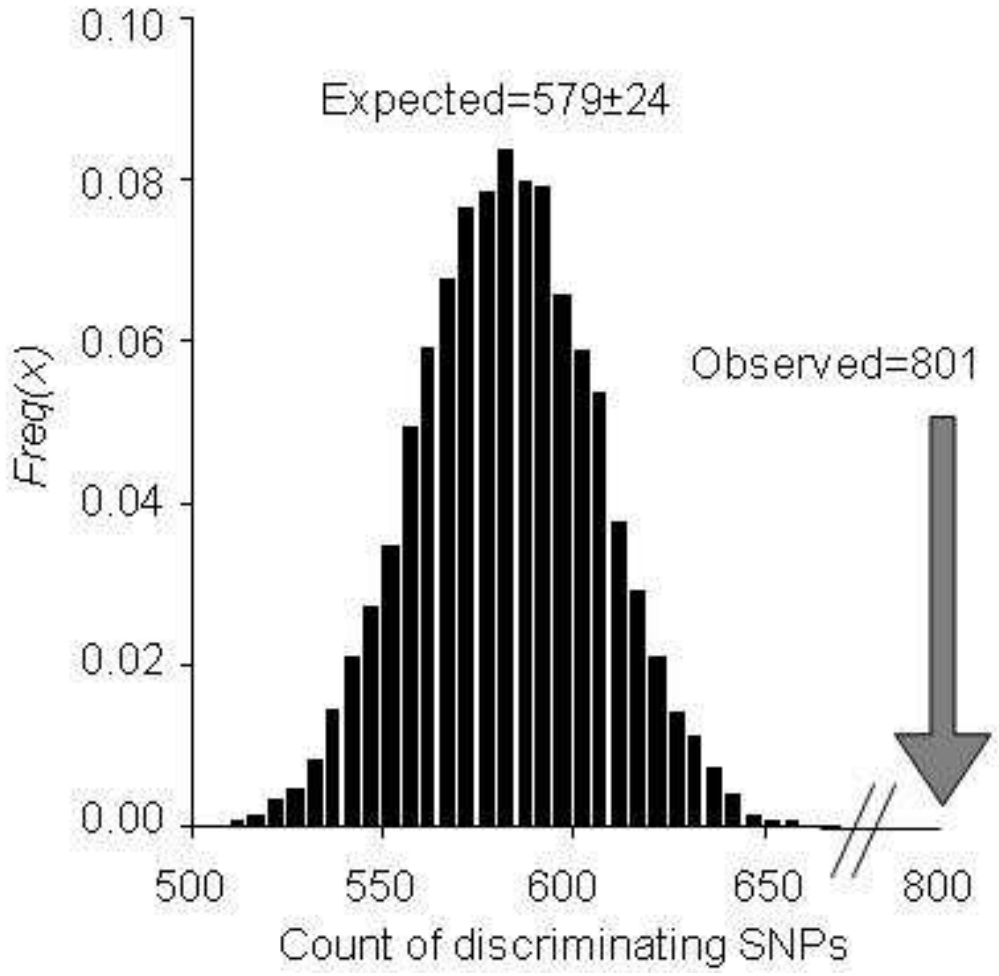


Figure 6.2. Expected versus observed number of discriminating SNPs. The *N. meningitidis* genome sequence alignment was simulated to yield a null distribution of the expected number of discriminating SNPs given the background variation. The expected distribution is compared to the observed number of discriminating SNPs.

SINGLE NUCLEOTIDE POLYMORPHISMS DISCRIMINATE BETWEEN DISEASE ASSOCIATED AND CARRIED GENOMES OF *N. MENINGITIDIS*

Previous comparative genomic sequence analyses of disease associated versus carried isolates of *N. meningitidis* failed to turn up evidence of obvious genomic differences between the two groups based on the presence or absence of any particular genes. In order to evaluate the genomic basis of the difference between disease associated and carried genomes here, we focused our analysis on differences at the level of individual nucleotide variation. To do this, we compared genomic sequences of 8 disease associated and 3 carried isolates of *N. meningitidis* (Table 6.1). Whole genome sequences of the *N. meningitidis* isolates were aligned as described in the Materials and Methods. There are 69 long orthologous regions (Local Collinear Blocks) conserved among all 11 genomes, and the total length of the *N. meningitidis* genome sequence alignment is 1,140,825 positions, 841,520 of which are absolutely conserved.

We identified aligned positions that show variation among genomes of *N. meningitidis* as single nucleotide polymorphisms (SNPs). These SNPs include positions with insertion/deletion (*i.e.* alignment gap) variation among genomes. There are a total of 299,305 SNPs in the whole genome *N. meningitidis* sequence alignment. We characterized SNPs that discriminate between the genomes of disease associated versus carried isolates of *N. meningitidi*s as those with mutually exclusive nucleotide patterns, including gap characters, between the two sets of sequences (Figure 6.1A). There are 801 such discriminating SNPs, and they are distributed across the entire *N. meningitidis* genome (Figure 6.1B). Discriminating SNPs were associated with individual *N. meningitidis* genes if they were found in the coding region or within 1 kb of a gene. The frequency distribution of discriminating SNPs per gene shows that the majority of SNP genes are associated with only one, or very few, discriminating SNPs (Figure 6.1C). Taken together, the genomic and frequency distributions of discriminating SNPs indicate that there is no systematic bias in how these SNPs are sampled in genomic and/or alignment space.

The disease associated versus carried isolate discriminating SNPs represent a catalog of individual nucleotide variation with potential implications for understanding the genomic basis of virulence in *N. meningitidis*. However, this study covers a limited set of genomes, and it is likely that when additional or different sets of genome sequences are compared, distinct sets of discriminating SNPs will be observed. Furthermore, with respect to the sequence data analyzed here, it is possible that we observe this number of discriminating SNPs (801) simply by chance alone given the large number of SNP positions found in the whole genome alignment analyzed here (299,305). We performed a simulation analysis to evaluate the probability of observing 801 discriminating SNP positions by chance alone given the background sequence variation among the aligned genomes. To do this, the isolate identities of the aligned genomes were randomly permuted 10,000 times, and for each permutation, a number of discriminating SNPs over the permuted alignment was computed. This procedure resulted in a null distribution of discriminating SNP counts, parameterized against the actual background variation, against which we compare our observed value (Figure 6.2). The value of the observed number of discriminating SNPs falls far outside of the range of the entire set of simulated values. Accordingly, the observed number of discriminating SNPs is significantly greater than can be expected by chance alone given the background variation among the genomes of *N. meningitidis* studied here (z=9.4, P=5.9e^{-21} z-test; or P<10e^{-4} based on the simulation).

DISCRIMINATING POLYMORPHISMS REFLECT PHENOTYPE RATHER THAN SHARED EVOLUTIONARY HISTORY.

It is clear that there are many more SNPs that discriminate between the disease associated versus carried genomes of *N. meningitidis* analyzed here than can be expected by chance alone. While these SNP data are suggestive of genomic differences with phenotypic relevance for virulence, they may also be attributed to shared evolutionary history. In other words, the abundance of discriminating SNPs may simply reflect the fact that the disease associated

isolates analyzed here are more closely related to each other than to the carried isolates and *vice versa*. If this is indeed the case, then the overall nucleotide sequence variation observed here should partition the disease associated versus carried isolates into two discrete groups of related genomes. We evaluated how *N. meningitidis* genome sequence variation is partitioned among the disease associated versus carried isolates studied here in several different ways: 1) using phylogenetic analyses to infer the evolutionary history of the genomes, 2) using a standard population genetic measure – the Fixation index (F_{ST}) – to evaluate how nucleotide variation is partitioned within and between the disease associated versus carried groups and 3) using naive Bayesian clustering of the observed SNP variation.

We reconstructed the phylogenies of the *N. meningitidis* genomes analyzed here in order to assess their evolutionary relationships. Specifically, we sought to evaluate whether the disease associated and carried genomes form distinct phylogenetic groups, each of which shares a unique common ancestor (*i.e.* monophyletic clades). To do this, we first aligned the seven house-keeping loci used for multi-locus sequence typing (MLST) (Holmes *et al.*, 1999; Maiden *et al.*, 1998) among the 11 genomes analyzed here and reconstructed a phylogeny based on the concatenated alignment (Figure 6.3A). The MLST sequence based phylogeny groups *N. meningitidis* genomes faithfully according to their sequence type (ST) and serogroup. For instance, all 3 ST-11 genomes (FAM18, M15141 & M9261) have absolutely identical MLST loci sequences, and the serogroup B genomes (M10699 & MC58) group together. The most important feature of this tree is the fact that the disease associated and carried isolates do not form separate and distinct monophyletic groups. In fact, disease associated and carriage genomes are grouped together on this tree with high bootstrap support. This finding indicates that the large number of disease associated versus carried isolate discriminating SNPs is not based on shared evolutionary history alone.

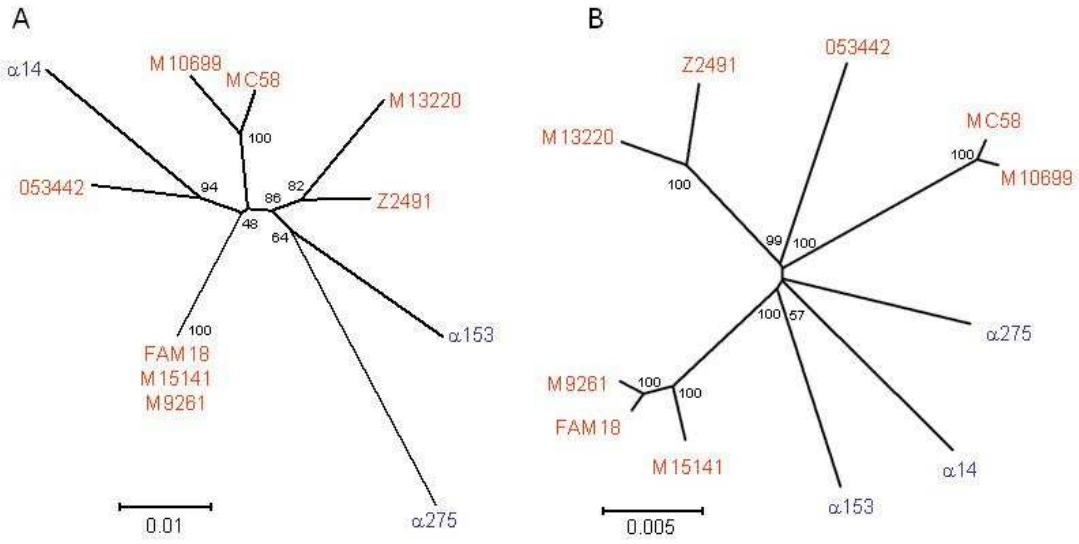


Figure 6.3. Phylogenetic analysis of disease associated and carried *N. meningitidis* isolate genomes. Disease associated genomes are shown in red and carried genomes are in blue. (A) A tree based on an alignment of the multilocus sequence typing loci: *abcZ*, *adk*, *aroE*, *fumC*, *gdh*, *pdhC*, and *pgm* (Holmes *et al.*, 1999; Maiden *et al.*, 1998). (B) A tree based on a whole genome alignment.

In attempt to gain more resolution for phylogenetic analysis, pairwise distances computed from the entire whole genome alignment were used (Figure 6.3B). This version of the phylogeny is slightly different with respect to some of the less supported internal branches, but disease associated and carried genomes still do not form distinct mutually exclusive evolutionary groups. On this tree, the carried genomes represent basal evolutionary lineages that are nested in between more derived lineages made up of disease associated genomes that are closely related to each other but distantly related to other disease associated isolates.

DISTRIBUTION OF SNP VARIATION AMONG *N. MENINGITIDIS* GENOMES

In addition to phylogenetic analysis, we directly evaluated how SNP variation is distributed within and between disease associated versus carried genomes using a population genetic measure – the Fixation index (F_{ST}). F_{ST} is a population differentiation measure that is based on polymorphism data; it measures the difference of between population variation from within population variation. High values of F_{ST} (close to 1) indicate that polymorphisms tend to segregate between rather than within groups and reveal highly differentiated populations. Using the SNP variation data in the whole genome sequence alignment, we measured F_{ST} taking the disease associated versus carried groups of genomes as two putative populations (Figure 6.4). F_{ST} for disease associated versus carriage genomes is low (0.047) and statistically indistinguishable from a null distribution of F_{ST} values calculated using a simulation procedure similar to that described for the discriminating SNP analysis (z=0.5, P=0.6 z-test; or P=0.3 based on the simulation). In other words, the polymorphism data based on the whole genome alignment do not provide evidence for population subdivision between disease associated and carried genomes of *N. meningitidis* as an explanation for the excess of observed discriminating SNPs.

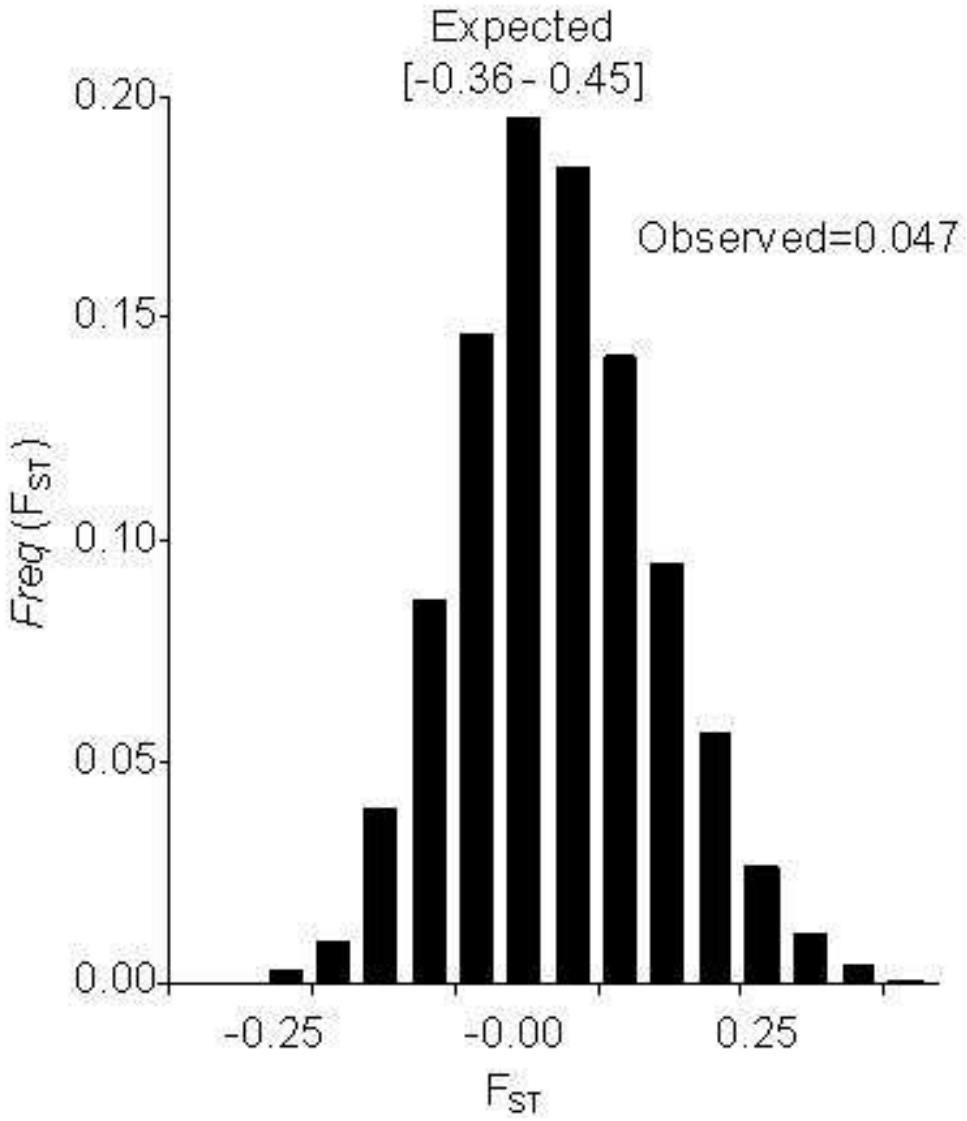


Figure 6.4. Differentiation of SNPs within and between grouped disease associated and carried *N. meningitidis* genomes based on the fixation index (FST). The *N. meningitidis* genome sequence alignment was simulated to yield a null distribution of the expected range of F_{ST} values. The expected F_{ST} range is compared to the observed value.

We also used a naïve Bayesian classification approach to partition SNP variation among the *N. meningitidis* genomes studied here using *K*-means clustering (Figure D.1). This approach was implemented with the program STRUCTURE in order to address two questions: 1) What is the optimal value of *K*? In other words how many genome groups do the SNP data indicate? and 2) Are disease associated versus carried genomes segregated into distinct groups based on the SNP data. Based on a user defined value of *K*, STRUCTURE assess the statistical likelihood of observing the data given *K* and assigns individual SNPs into each group. For any given genome, the fraction of SNPs in each group can then be ascertained. This allows for a determination of the extent to which a given genome faithfully maps to one group or the other. The optimal value of *K* according to this analysis is 3, not 2 as may be expected if disease associated and carried genomes formed distinct groups (Figure D.1A). In addition, the SNP variation at *K*=3 does not cleanly partition among individual genomes (Figure D.1B). This result is consistent with a high level of recombination among *N. meningitidis* strains and is indicative of reticulate evolution and/or shared polymorphisms. This is particularly true for the carried isolate genomes, which appear to have the most mixed ancestry in terms of the three SNP clusters. Disease associated genomes are less hybrid in general with respect to SNP polymorphism, and there are 5 disease associated isolates with SNPs that segregate almost perfectly into 1 of 2 clusters. Apparently, there have been abundant opportunities for genetic exchange subsequent to the divergence of these genomes and specific nucleotide variants acquired via recombination may be important markers for virulence.

ASSOCIATION OF DISCRIMINATING SNPS WITH *N. MENINGITIDIS* GENES

Disease associated versus carried genome discriminating SNPs were associated with genes if they were found either within or proximal to coding sequences (Table 6.3). A total of 527 discriminating SNPs were associated with 113 *N. meningitidis* genes – hereafter referred to as

'SNP genes'. One-hundred eighty-three of these discriminating SNPs map to gene-proximal non-coding sequences, and 344 map to coding sequences.

The 344 coding sequence discriminating SNPs were classified as either non-synonymous or synonymous based on whether or not they correspond to differences in encoded amino acid sequences between disease associated versus carried genomes. There are 319 (92.7%) non-synonymous discriminating SNPs compared to only 25 (7.3%) synonymous SNPs. Since ~74.3% of all possible coding sequence substitutions are non-synonymous (Nei & Gojobori, 1986), these data indicate a statistically significant excess of non-synonymous substitutions among the coding sequence discriminating SNPs (χ^2=60.6, *df*=1, *P*=7e-15). The fact that coding sequence discriminating SNPs are more likely to result in changes at the level of protein sequence is consistent with the notion that these positions contribute to phenotypic differences between disease associated versus carried genomes. Below, we explore the possible functional implications of discriminating SNPs in more detail.

TABLE 6.3. Genomic features of discriminating SNPs.

Discriminating SNP class	Count
All	**801**
Non-genic	**274**
Gene associated	**527**
Non-coding gene associated	**183**
Coding sequence gene associated	**344**
Non-synonymous coding sequence	**319**
Synonymous coding sequence	**25**

FUNCTIONAL ANALYSIS OF *N. MENINGITIDIS* DISCRIMINATING SNP GENES

The SNP genes (proteins) were evaluated with respect to a wide variety of functional characteristics to determine if they are enriched for any particular functions or features that may be related to virulence. The SNP genes were not found to be enriched for function at the cell periphery (*i.e.* signal peptides or lipoproteins), horizontally transferred genes or putative virulence factors. However, the SNP genes were found to be enriched for 48 specific gene ontology (GO) functional annotations (Table 6.4). Many of the enriched SNP genes span multiple functions due to the GO hierarchy; there are a total of 45 SNP genes among the overrepresented functional categories (Table 6.5). Of the 45 genes with overrepresented functions, 10 are most closely related to virulence and we explore the potential functional relevance of these SNP genes below. These 10 SNP genes are grouped into three distinct categories: “oxidoreductases,” “cell wall,” and “immune system evasion and other proteases” and represent a potential list of prioritized targets for future experimental interrogation based on the initial genome comparisons done here.

TABLE 6.4. Overrepresented functions and categories in SNP genes as compared to all genes in M13220.

Categories of SNP genes were analyzed using the hypergeometric distribution against the background of all genes.

Category[a]	Description	Number of genes in the genome[b]	SNP gene count[c]	*P*[d]
GO:0006596	polyamine biosynthetic process	2	2	**0.0028**
GO:0006595	polyamine metabolic process	2	2	**0.0028**
GO:0044403	symbiosis, encompassing mutualism through parasitism	2	2	**0.0028**
GO:0009405	pathogenesis	2	2	**0.0028**
GO:0044419	interspecies interaction between organisms	2	2	**0.0028**
GO:0009065	glutamine family amino acid catabolic process	2	2	**0.0028**
GO:0006575	cellular amino acid derivative metabolic process	14	4	**0.0052**
GO:0006519	cellular amino acid and derivative metabolic process	126	14	**0.0061**
GO:0006576	biogenic amine metabolic process	8	3	**0.0069**
GO:0045454	cell redox homeostasis	18	4	**0.0134**
GO:0042398	cellular amino acid derivative biosynthetic process	10	3	**0.0137**
GO:0019725	cellular homeostasis	20	4	**0.0195**
GO:0006520	cellular amino acid metabolic process	118	12	**0.0213**
GO:0044106	cellular amine metabolic process	118	12	**0.0213**
GO:0042401	biogenic amine biosynthetic process	5	2	**0.0256**
GO:0016744	transferase activity, transferring aldehyde or ketonic groups	5	2	**0.0256**
GO:0051704	multi-organism process	5	2	**0.0256**
GO:0009308	amine metabolic process	137	13	**0.0283**
GO:0009056	catabolic process	96	10	**0.0299**
GO:0043436	oxoacid metabolic process	153	14	**0.0307**
GO:0019752	carboxylic acid metabolic process	153	14	**0.0307**
GO:0030170	pyridoxal phosphate binding	23	4	**0.0314**
GO:0005496	steroid binding	23	4	**0.0314**
GO:0070279	vitamin B6 binding	23	4	**0.0314**
GO:0006082	organic acid metabolic process	156	14	**0.0355**
GO:0008289	**lipid binding**	**24**	**4**	**0.0362**

Table 6.4 (continued)

GO:0009063	**cellular amino acid catabolic process**	**6**	**2**	**0.0371**
GO:0046395	carboxylic acid catabolic process	6	2	**0.0371**
GO:0016054	organic acid catabolic process	6	2	**0.0371**
GO:0009310	amine catabolic process	6	2	**0.0371**
GO:0042180	cellular ketone metabolic process	157	14	**0.0373**
GO:0042592	homeostatic process	25	4	**0.0414**
GO:0004527	exonuclease activity	15	3	**0.0426**
GO:0009057	macromolecule catabolic process	75	8	**0.0443**
GO:0044248	cellular catabolic process	38	5	**0.0494**
GO:0016796	exonuclease activity, active with either ribo- or deoxyribonucleic acids and producing 5'-phosphomonoesters	7	2	**0.0501**
GO:0043648	dicarboxylic acid metabolic process	16	3	**0.0504**
GO:0043603	cellular amide metabolic process	16	3	**0.0504**
GO:0032559	adenyl ribonucleotide binding	165	14	**0.0534**
GO:0005524	ATP binding	165	14	**0.0534**
GO:0016667	oxidoreductase activity, acting on sulfur group of donors	17	3	**0.0589**
GO:0033554	cellular response to stress	40	5	**0.0595**
GO:0030554	adenyl nucleotide binding	184	15	**0.0620**
GO:0001883	purine nucleoside binding	184	15	**0.0620**
GO:0001882	nucleoside binding	185	15	**0.0644**
GO:0044270	nitrogen compound catabolic process	8	2	**0.0645**
GO:0051716	cellular response to stimulus	41	5	**0.0650**
GO:0004175	**endopeptidase activity**	**29**	**4**	**0.0660**

[a] Category, as determined by comparing with GO

[b] Number of genes in the genome belonging to this category (2108 total genes in the genome)

[c] Number of SNP genes belonging to this category (113 total SNP genes)

[d] *P*-value for the hypergeometric testing for enrichment of the GO category

TABLE 6.5. SNP genes from virulence-related over-represented gene categories.

Name[a]	Locus Tags[b]	functional annotation	No. SNPs[c]	Inter-[d]	Intra-[e]	Nsyn.[f]	Syn.[g]
Oxidoreductases							
dsbA1	NMA2209 NMB0278 NMC0273	putative thiol:disulphide interchange protein	1	1	0	0	**0**
	NMA0253 NMB0006 NMC2144	putative periplasmic thiredoxin	1	1	0	0	**0**
Immune System Evasion and Other Proteases							
iga	NMA0905 NMB0700 NMC0651	IgA1 protease	7	0	7	1	**6**
lon	NMA1398 NMB1231 NMC1131	putative ATP-dependent protease	15	0	15	4	**11**
prlC	NMA0054 NMB0214 NMC0206	oligopeptidase A	1	1	0	0	**0**
	NMA2172 NMB0315 NMC1856	putative metallopeptidase	5	0	5	1	**4**
	NMA1066 NMB0855 NMC0795	putative periplasmic protein	17	15	2	1.5[h]	**0.5**[h]
nhhA	NMA1200 NMB0992 NMC0978	putative surface fibril protein	31	30	1	1	**0**
Cell Wall							
mviN	NMA2210 NMB0277 NMC0272	putative inner membrane protein	6	0	6	4	**2**
ftsE	**NMA0254 NMB0007 NMC2145**	**ABC transporter ATP-binding protein**	**1**	**1**	**0**	**0**	**0**

[a] If it is defined, the consensus gene name of the homolog found in strains Z2491/MC58/FAM18

[b] If defined, the locus tag of the homologous gene in strains Z2491/MC58/FAM18

[c] Total number of gene associated SNPs

[d] Number of SNPs found flanking the coding sequence of the gene

[e] Number of SNPs found within the coding sequence

[f] Number of nonsynonymous SNPs within the coding sequence

[g] Number of synonymous SNPs within the coding sequence

[h] At one discriminating SNP position, a substitution resulted in a nonsynonymous substitution in 50% of the discriminating pattern.

Oxidoreductases. The locus with homology to NMA0253/NMB0006/NMC2144 codes for a thioreductase with homology to the *tlpA* gene, which was characterized in *N. gonorrhoeae* (Achard *et al.*, 2009). While it is possible that TlpA homolog could play a role in pilin disulfide bond reduction, it has been speculated that it has a role in ameliorating oxidative stress from lysosomes and therefore promotes intracellular survival. Another SNP gene product, DsbA1, is a lipoprotein that oxidizes sulfide bonds to create disulfide bonds, and it gets its oxidizing power from the electron transport chain (Tinsley *et al.*, 2004). Incidentally, there are several other SNP genes implicated in the pathways around the electron transport chain including those with the GO accession GO:0016651 whose function is "oxidoreductase activity, acting on NADH or NADPH" (File D.1).

That both *dsbA* and the *tlpA* homolog are SNP genes with overrepresented functions may indicate that disulfide bonds are highly regulated in disease associated genomes of *N. meningitidis*. High regulation would be especially pertinent given that certain pili require disulfide bond formation for their functions, especially PilE and PilQ, both of which are involved in the formation of type IV pili (Tinsley *et al.*, 2004; Vivian *et al.*, 2009). Type IV pili are essential for intimate adhesion of capsulated meningococci to epithelial cells (Nassif *et al.*, 1997). At least one other SNP gene, *pglA*, codes for a protein with a function that may affect pili as well. PglA is responsible for glycosylating a disaccharide on pilin, specifically at a serine residue (Power & Jennings, 2003). At this site, PglA adds galactose (α1,3) to the disaccharide structure. As a consequence of glycosylation, adherence to epithelial cells is debilitated (Marceau *et al.*, 1998). Therefore one difference between disease associated and carried isolates could be variation in adherence to endothelial cells. Consistent with this prediction, Joseph *et al* recently showed that adhesins are up-regulated in an invasive isolate of *N. meningitidis* (MC58) compared to a closely related carried isolate (α710) (Joseph *et al.*, 2010).

Cell Wall. Although *mviN* is uncharacterized and is a putative gene in meningococcus, its products have been experimentally characterized in *Escherichia coli* and *Salmonella typhimurium* which are also gram-negative (Benjamin *et al.*, 1991; Inoue *et al.*, 2008). A mutation in the functional *S. typhimurium* homolog of *mviN* renders an otherwise avirulent isolate virulent. In *E. coli*, *mviN* has been shown to be essential for murein synthesis.

Although *ftsE* is not directly involved in virulence, it is a transporter upstream of *bolA* which has been characterized in *E. coli* (Aldea *et al.*, 1988). Because it is only 61 nucleotides away from its neighbor and has the same transcriptional orientation, the *bolA* homolog is likely in the same operon as *ftsE* and under the same transcriptional regulation (Price *et al.*, 2005). BolA has been found to contribute to retaining normal cell morphology during stationary phase and in conditions of starvation (Santos *et al.*, 2002). Moreover, *bolA* regulates two penicillin binding proteins (PBPs) (PBP5 and PBP6). PBPs are involved in the synthesis of murein and can be targets of β-lactam antibiotics. Gene expression of *bolA* is more significant during exponential growth phase than stationary phase, which supports a hypothesized role – that *bolA* is critical for adaptation of cell morphology and cell wall composition to the growth conditions (Santos *et al.*, 2002).

Immune System Evasion and Other Proteases. NhhA has a few purported functions including evading complement deposition and the resulting formation of the membrane attack complex (MAC) (Sjolinder *et al.*, 2008) as well as autotransporter and adhesin activity (Scarselli *et al.*, 2006). Δ*nhhA* mutants have reduced adherence to host cells and have more MAC deposition. Therefore NhhA is an adhesin and it contributes to meningococcal immune evasion.

IgA1 protease (*iga*) was classically thought to cleave only IgA antibodies which are secreted at mucosal surfaces (*e.g.,* the nasopharynx where meningococci can live) (Mistry & Stockley, 2006). By cleaving Fabα from Fcα fragment of an IgA antibody, meningococcal opsonization is

limited, thereby decreasing agglutination and opsonophagocytosis. Furthermore after IgA cleavage, the meningococcal cell surface can bind the remanant "self" Fabα fragments to mask the pathogen from the host immune system.

While it still holds true that it cleaves IgA antibodies, IgA1 protease has other cleavage targets such as human chorionic gonadotropin, granulocyte-macrophage colony stimulating factor, the CD8 surface antigen of cytotoxic T lymphocytes, and LAMP1 (Mistry & Stockley, 2006). LAMP1 is a protein with a hypothetical structural function in lysosomes. Not only has LAMP1 cleavage by IgA1 protease been shown to be a major factor in meningococcal pathogenesis (Hauck & Meyer, 1997; Lin *et al.*, 1997), but some anecdotal evidence shows that LAMP1 cleavage helps the meningococcus escape from the phagosome (Hauck & Meyer, 1997).

The gene *lon*, coding for Lon protease, has not been characterized in *Neisseria* but has been characterized in several other species including *E. coli* and *Streptococcus pneumoniae* (Ingmer & Brøndsted, 2009). While a common cleavage sequence motif has not been found in its several targets, Lon is thought to target particular tertiary sequences. These targets are likely to be in a nonglobular conformation with exposed hydrophobic patches or other specific structural motifs (Tsilibaris *et al.*, 2006). Lon permits intracellular survival of *S. typhimurim* through down-regulation of *Salmonella* Pathogenicity Island I (SP-I) by cleaving an activator of SP-I (Boddicker & Jones, 2004). Furthermore, Lon might increase resistance to oxidative stress especially in the cases of respiratory bursts from lymphocytes and also might increase resistance to low phagosomal pH (Takaya *et al.*, 2003). It is interesting to note that the gene with homology to oxidoreductase TlpA has the same function in increasing resistance to low phagosomal pH.

The gene *prlC* is a homolog of *opdA*, found in *E. coli* and *Salmonella enterica* serovar Typhimurium, and is a metalloprotease (Conlin & Miller, 2000). Not much is known about *prlC*, but it is located in a σ^{32}-dependent heat shock operon which makes *prlC* a putative heat shock

protein (HSP). Many HSPs are induced by other environmental changes such as interaction with a eukaryotic host (Du *et al.*, 2005). Therefore *prlC* might aid in stress that is associated with interaction of the human host.

DYNAMICS OF *N. MENINGITIDIS* COLONIZATION, CARRIAGE AND DISEASE

The *N. meningitidis* genomes analyzed here were chosen based on the fact that the isolates originated from either diseased individuals, what we refer to as disease associated, or asymptomatic carriers, referred to here as carried isolates. Thus, what we have observed in this analysis is essentially a snap-shot in time and place of a set of particular nucleotide variants that distinguish one group of *N. meningitidis* disease associated isolates from a group of carried isolates. It must be noted, however, that in any individual, an invasive meningococcal disease case originates from an asymptomatic colonization state (Jolley *et al.*, 2005; Meyers *et al.*, 2003). Transition of the bacterium from a colonization to disease causing state may be accompanied by the acquisition of specific nucleotide variants; it may be dependent on the genetic background of the human host (Davila *et al.*) or other environmental factors, or it may be caused by some combination of these factors. Thus, it is formally possible that the carried isolate genomes studied here could evolve rapidly to become invasive genomes that cause disease. Indeed, we have shown here that disease associated and carried genomes do not form mutually exclusive evolutionary groups, consistent with repeated changes between these states over time (Figures 6.3 and 6.4). Even closer evolutionary relationships between disease associated and carried isolates of *N. meningitidis* have been demonstrated elsewhere (Jolley *et al.*, 2005). Further consistent with the possibility of changing between disease associated and carried states, two out of the three carried isolates investigated here are members of a strain lineage that has been observed to cause a low percentage of the invasive disease in the geographic area from which it was obtained (Schoen & Claus, 2006). The dynamics of all these

factors necessitate that the discriminating SNPs and SNP genes identified here be treated with caution, since they could easily change depending on changes in the disease causing potential of the genomes in which they are found.

CONCLUSION

Because the presence or absence of genes alone does not determine meningococcal virulence (Bille *et al.*, 2005; Hotopp *et al.*, 2006; Perrin *et al.*, 2002; Schoen *et al.*, 2008; Snyder & Saunders, 2006; Stabler *et al.*, 2005), we sought smaller differences in the form of SNPs between the genomes of 8 disease associated and 3 carried isolates of *N. meningitidis*. We identified 801 discriminating SNPs and the genes associated with 527 of them. Of the 113 SNP genes identified, functional analysis indicates 10 as the most likely targets of further investigation based on their possible roles in virulence.

In addition to the caveats described previously, it should be noted that the analyses performed here are limited by the relatively small number of complete genome sequences that were analyzed: 8 disease associated and 3 carried isolates. Consequently, the particular set of discriminating SNPs characterized here, along with the list of SNP genes, will likely change as additional genome sequences are characterized and compared. Thus, a more definitive understanding of genome-level differences between disease associated and carriage isolates of *N. meningitidis* will require the analysis of additional genomes.

ACKNOWLEDGEMENTS

This work was supported in part by Centers for Disease Control and Prevention (1 R36 GD 000075-1) to L.S.K.; Bioinformatics program, Georgia Institute of Technology to N.V.S.; Defense Advanced Research Projects Agency (HR0011-05-1-0057) to A.O.K.; Georgia Research Alliance (GRA.VAC09.O) to B.H.H., I.K.J. and L.W.M.; and The Alfred P. Sloan Foundation (BR-4839) to I.K.J.

We wish to thank the Meningitis and Vaccine Preventable Diseases Branch at CDC, especially Nancy Messonnier and Tom Clark for helpful discussions. We also wish to thank Ahsan Huda for helpful discussions.

The authors also wish to thank each organization or individual who provided some of the invasive isolates in this study. We thank the Active Bacterial Core Surveillance Team (M10699), the Emerging Infection Programs Network (M15141), RITM (M9261), and Philippines (M13220).

CHAPTER 7

THE GENOMIC BASIS OF A NONGROUPABLE *NEISSERIA MENINGITIDIS* ISOLATE

ABSTRACT

Neisseria meningitidis is a gram-negative bacterium that has the potential to cause acute diseases including meningitis and septicemia, but its spread can be reduced when its polysaccharide capsule type is inferred. Its capsule can be classified into 12 different serogroups determined from the chemical composition of the capsule. There is an antibody assay to infer the type of polysaccharide capsule and ideally, this assay will return exactly one positive result. However occasionally there are no or multiple reactions, i.e. nonagglutinating or polyagglutinating, respectively. In these cases, the isolate is characterized as "nongroupable" (NG). We have received and characterized an isolate (M16917) as NG-polyagglutinating. To understand the origins of M16917, we performed multilocus sequence typing and whole genome profiling, and to better understand the origin of the capsule, we created a phylogeny of the capsule polymerase gene. The results of these analyses are that the genome has its origins in serogroup C while the capsule is most related to serogroup B, thus strongly supporting a capsule switching event. A three-way multiple sequence alignment between M16917 and representative serogroup B and C genomes was constructed and analyzed to uncover likely recombination breakpoints. Four breakpoints were uncovered which flank two different recombination events. These two cassettes are likely of serogroup B origin. One breakpoint separates the capsule polymerase gene into serogroup B origin and serogroup C origin. Previous studies have indicated that it may be possible that 1) a recombination event in the middle of the

capsule polymerase gene can alter its linkage specificity and 2) that certain mutations may cause a hybrid of two capsules. Due to the recombination event in the M16917 capsule polymerase gene and in light of previous studies, it is possible that M16917 has a chimeric B/C polymerase. Future experimental characterization of the capsule structure could confirm this hypothesis.

INTRODUCTION

Neisseria meningitidis is a gram-negative bacterium that has the potential to cause sudden and acute diseases including meningitis and septicemia (Rosenstein *et al.*, 2001). It is highly competent and can exchange DNA through homologous recombination especially within its genus (Holmes *et al.*, 1999; Jolley *et al.*, 2005). *N. meningitidis* is protected from host-mediated, complement-dependent bacteriolysis and phagocytosis by a polysaccharide capsule (Rosenstein *et al.*, 2001) and is classified into 12 serogroups determined from the chemical composition of the capsule. Distinct serogroup capsule compositions differ by either their particular sugar makeup or the linkages between sugars (Frosch & Vogel, 2006; Popovic & Ajello, 2003). The most typically disease-associated serogroups are A, B, C, W135, and Y (Yazdankhah *et al.*, 2004). Effective vaccines are available for the disease-associated serogroups A, C, W135, and Y (Bilukha & Rosenstein, 2005), whereas there is no protective vaccine specifically against the serogroup B capsule (Harrison *et al.*, 2009).

During an outbreak of meningococcal disease, determination of the particular serogroup provides information to public health officials as to whether or not vaccination can be used as a prevention measure. Typically, if the serogroup of isolates from infected individuals can be characterized as one for which a vaccine exists, then the vaccine will be used prophylactically among close contacts of those individuals to help contain the outbreak. This approach will only work if there is an effective vaccine available for the characterized serogroup. For these

reasons, rapid identification of *N. meningitidis* capsule serogroups is a critical component of public health response to meningococcal disease.

The chemical composition of the *N. meningitidis* polysaccharide capsule is classically inferred using the slide agglutination serogrouping assay (SASG) (Popovic *et al.*, 1999; Popovic & Ajello, 2003). The SASG assay reveals the capsule type with serogroup-specific antibodies that yield positive or negative reactions. Ideally, exactly one reactivity will be positive per isolate, but occasionally there are multiple or no positive reactions. In these cases, the isolate is characterized as "nongroupable" (NG). Isolates that have multiple positive SASG results are referred to as NG-polyagglutinating. Mothershed *et al.* introduced the serogroup-specific real-time PCR assays (SGS-PCR) to help resolve NG SASG results (Mothershed *et al.*, 2004). The SGS-PCR assay is a real-time PCR test using primers and probes specific to each serogroup. The gene target for these assays is typically the gene that encodes capsule polymerase specific to each serogroup. The CDC Meningitis Laboratory and other reference labs routinely perform both SASG and SGS-PCR to group isolates quickly and unambiguously.

Despite the availability of these two assays for the characterization of *N. meningitidis* capsule serogroups, there remain instances of NG isolates and at times the SASG and SGS-PCR results do not agree. Examination these NG isolates reveals three distinct classes: autoagglutinating, nonaggultinating and polyagglutinating. Autoagglutinating isolates agglutinate in the saline control for the SASG assay via a mechanism unrelated to serogroup specificity. Non-agglutinating isolates show no evidence of capsule expression and a number of such isolates have been genetically characterized. This work showed that non-aggultination can result from a loss of capsule expression through point mutations and/or phase variation as well as partial or complete deletions of capsule operon sequences (Dolan-Livengood *et al.*, 2003; Weber *et al.*, 2006). Polyaggutinating isolates are defined as showing cross-reactivity to two or

more serogroup specific antisera. At this time, the genetic basis of polyaggultination is less well understood. Polyagglutination has been hypothesized to result from the expression of multiple capsule types, but has more often been attributed simply to the quality of the antisera reagents used in the SASG assay.

In this report, we sought to characterize the genomic basis of an NG polyaggulitination result for a cerebrospinal fluid (CSF) isolate (M16917) taken from a patient with acute meningococcal disease. To this end, we characterized and analyzed its complete genome sequence. The genome was characterized using 454 pyrosequencing and the sequence was assembled and analyzed using the CG-Pipeline genome analysis software tool. The M16917 isolate was originally characterized as polyagglutinating based on the SASG assay and as serogroup B based on SGS-PCR. Multi-locus sequencing typing (MLST) analysis of the M16917 genome sequence as well as whole-genome sequence comparisons among multiple completely sequences strains of *N. meningitidis* revealed M16917 to be a member of the ST-11 sequence group despite the fact that ST-11 strains are typically associated with serogroup C capsule. Subsequent examination of the capsule locus revealed sequences with homology to both serogroup B and serogroup C strains of *N. meningitidis*. Comparison with capsule loci from completely sequenced *N. meningitidis* B and C serogroup reference strains was used to reveal a complex chimeric B/C structure for the M16917 capsule locus by defining the strain origins of adjacent capsule locus subsequences along with the specific locations of recombination breakpoints. The M16917 capsule locus results from two separate B/C recombination events and interestingly the capsule polymerase gene itself is a chimeric sequence containing both B and C type sequences. These results underscore the complexity of genomic rearrangements that can occur at the capsule locus and suggest a specific genomic basis for the NG results from the SASG and SGS-PCR assays of the M16917 strain.

EXPERIMENTAL PROCEDURES

ISOLATION AND CHARACTERIZATION OF M16917

The M16917 strain of *N. meningitidis* was isolated from the CSF of a 31 year old male with acute meningococcal disease. M16917 was initially characterized with the SASG assay by the Illinois Department of Public Health in Chicago, Illinois. M16917 was subsequently characterized by the CDC Meningitis Laboratory using both the SASG and SGS-PCR assays as previously described (Mothershed *et al.*, 2004; Popovic & Ajello, 2003). The genome sequence of M16917 was characterized with 454 pyrosequencing by the CDC core facility as previously reported (Kislyuk *et al.*, 2010). To clarify a homopolymeric region in the *adk* locus, we performed Sanger sequencing as described at pubmlst.org (Jolley *et al.*, 2004).

GENOME BASED MLST ANALYSIS

We employed a multilocus sequence typing (MLST) analysis, which characterizes bacterial isolates based on the alleles of seven highly conserved housekeeping genes (Holmes *et al.*, 1999; Maiden *et al.*, 1998), on the complete genome sequence of M16917. The genome sequence was assembled *de novo* using the CG-Pipeline (v0.2.2) software tool (Kislyuk *et al.*, 2010), and the assembly was compared to the PubMLST database of alleles for each of the seven housekeeping loci (Jolley *et al.*, 2004) using BLAST (v2.2.17)(Altschul *et al.*, 1997). We determined the presence of an allele based on whether the best hit of the BLAST analysis had 100 percent identity, with no gaps or mismatches, to that allele as well as an alignment length equal to the full length of the allele. The concatenation of all seven alleles yields a sequence type (ST) which ordinarily belongs to a family of STs called a clonal complex (CC).

REFERENCE BASED ASSEMBLY

We performed a reference based assembly of M16917 using Newbler (v2.0.00.20) against the previously characterized genomes: FAM18 (ST-11/serogroup C), MC58 (ST-74/serogroup B),

Z2492 (ST-4/serogroup A), and α275 (ST-22/serogroup W135) (Table 7.1). The quality of each reference based assembly was evaluated based on a number of metrics: the number of large contigs produced, the number of bases assembled, N50 size, and average contig size (See Table 7.2 for definitions). The assembly with the best overall metrics was used for further analysis in this study.

Table 7.1. *N. meningitidis* reference genomes used in this study.

Name	Serogroup	ST[a]	CC[b]	Accession
FAM18	C	ST-11	ST-11	AM421808.1
MC58	B	ST-74	ST-32	AE002098
α275	W135	ST-22	ST-22	AM889138
Z2491	A	ST-4	ST-4	AL157959

[a]Sequence Type

[b]Clonal Complex

WHOLE GENOME PROFILING

The relative quality scores from M16917 reference assembly against the previously characterized FAM18, MC58, Z2491 and α275 genomes were used in whole genome profiling to evaluate the most likely origin of the M16917 genome. For each individual nucleotide position of the reference genome in these assemblies, a cumulative quality score taken from the Newbler output was recorded, and positions that were not mapped by any M16917 reads were counted as gaps with a score of 0. These scores were averaged for non-overlapping 1000bp windows across the reference genomes. The distributions of reference assembly average scores along the genomes were visualized using color scale (Figure 7.1).

Table 7.2. Mapping results. A mapping of all 454 pyrosequencing reads of M16917 against these reference genomes produced several assembly statistics. The reference genome with the best statistics is more closely related to the query genome M16917. The best statistics are defined by a higher average quality, lower number of contigs, more bases mapped, larger average contig size, and larger N50. The metrics with FAM18 are consistently better than other mapping results.

	Average Quality[a]	# of Large Contigs[b]	Bases Mapped[c]	Average Contig Size	N50[d]
Z2491 (A)	56.1±16.2	392	1.93 Mb	4969	8671
MC58 (B)	54.3±17.8	428	1.94 Mb	4740	8713
FAM18 (C)	58.4±15.1	179	2.09 Mb	12176	31630
α275 (W135)	54.4±17.8	457	1.92 Mb	4160	6714

[a]The quality is given as a Phred score, q=-10log(p) where p is the probability of an incorrect base call.

[b]Large contigs are those greater than 500 bp.

[c]Mapped bases are those that were able to be aligned against the reference genome.

[d]N50 is a standard quality metric for genome assemblies that summarizes the length distribution of contigs. It represents the size N such that 50% of the genome is contained in contigs of size N or greater.

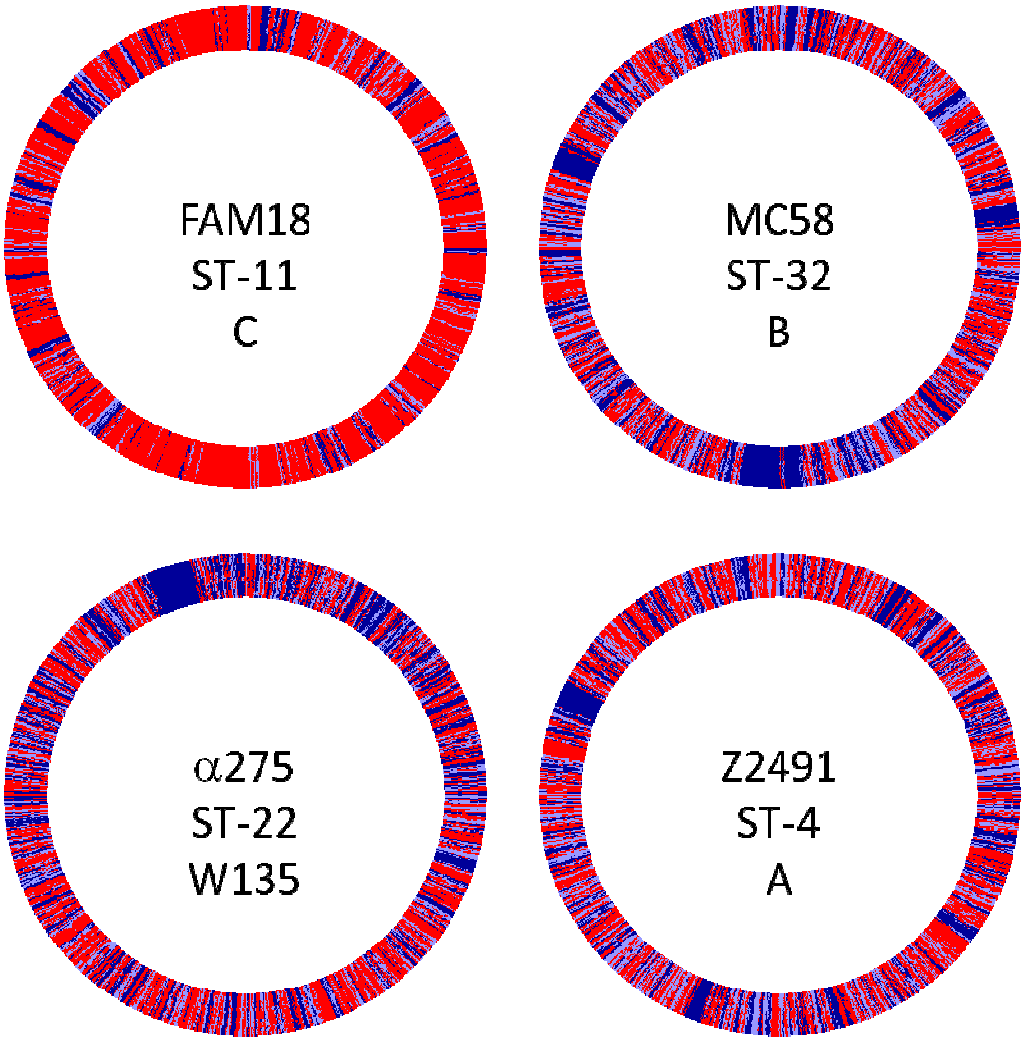


Figure 7.1. The M16917 trace data were mapped to each reference genome using Newbler. For every 1000 bases, cumulative quality scores of every base that mapped to the reference assembly were calculated. These cumulative quality scores were averaged together among the 1000 bases. Positions where no base mapped to the reference genome were given a quality score of zero. Red scores indicate high quality mapping, and blue scores indicate low quality mapping. Mapping against FAM18 shows the highest cumulative average quality scores, while MC58 shows the lowest.

IDENTIFICATION OF THE CAPSULE LOCUS AND CAPSULE POLYMERASE GENE

M16917 capsule locus sequence containing contigs were assembled *de novo* to avoid any biases that may be introduced by reference based assembly to one serogroup strain or another. *De novo* assembled contigs were then aligned to the *N. meningitidis* reference genomes MC58 (serogroup B) and FAM18 (serogroup C) using the program wgVISTA (Frazer *et al.*, 2004; Mayor *et al.*, 2000). Contigs that were visually confirmed to align to the capsule locus in either MC58 or FAM18 were retained for subsequent analysis. The identity of specific capsule locus genes, including the capsule polymerase gene used in the SGS-PCR assays, in these contigs was confirmed using BLAST.

PHYLOGENETIC ANALYSIS

Using MEGA (Tamura *et al.*, 2007) we performed a multiple sequence alignment (MSA) and subsequent neighbor-joining tree of reference capsule polymerase genes (FAM18, MC58, α275, and Z2491) and the nongroupable M16917 corresponding putative region.

IDENTIFICATION OF RECOMBINATION BREAK POINTS

We conducted a three way MSA of the capsular regions spanning *rfbA* to *tex* (the boundaries of the capsule locus) using CLUSTALW 2.0.10 (Larkin *et al.*, 2007). Within the alignment, we defined an informative site as any polymorphic site in which the nucleotide on M16917 matches either that of MC58 or FAM18, but not both. If the site on M16917 matched that of MC58 or FAM18, the position would be labeled as B or C, respectively. Those polymorphic sites that did not meet this criterion were not considered informative. A string of informative sites corresponding to serogroup classification was generated and is referred to as the informative string.

We employed a sliding window fifty positions long across the informative string and constructed a 2x2 contingency table splitting the window in two equal halves of twenty-five

nucleotides each. Pseudocounts equal to one were used to initiate each cell in the contingency table to facilitate χ^2 analysis. Analyses were performed for each contingency table to yield a *p* value, which was then transformed to a –log(*p*) value. In this analysis, a local maxima in the –log(*p*) graph indicates a possible recombination breakpoint.

RESULTS AND DISCUSSION

M16917 ISOLATION AND CHARACTERIZATION

In 2007, the Bacteriology Laboratory at The Illinois Department of Public Health, Chicago received an isolate of *N. meningitidis* cultured from the CSF of a 31 year old male with meningitis. They characterized the isolate as NG polyagglutinating using the SASG assay and sent it to CDC Meningitis Laboratory, which designated it as M16917. CDC performed the SASG and SGS-PCR assays, which resulted in NG polyagglutinating and serogroup B results, respectively.

We characterized and analyzed the complete genome sequence of M16917 in an effort to try and better understand the genomic basis of the NG polyagglutination phenotype. The genome of M16917 was sequenced using 454 pyrosequencing and assembled *de novo* using the CG-Pipeline genome analysis software tool (Kislyuk *et al.*, 2010). This assembly yielded 71 contigs >500bp in length with a longest contig of 192,528bp and an N50 value of 53,150bp. The total length of the genome was found to be 2,158,678bp and there are 2,337 predicted protein coding genes. These results are consistent with a number of previous *N. meningitidis* genome sequence projects (Kislyuk *et al.*, 2010).

ORIGINS OF THE M16917 GENOME

The origin of the M16917 genome sequence was evaluated using two approaches: a genome based MLST analysis and comparative whole genome profiling against *N. meningitidis* reference genomes from different serogroups. The seven loci used in MLST were located in the

M16917 assembly using BLAST, and the sequences for these loci were then queried against the PubMLST database (Jolley et al., 2004). This analysis resulted in an MLST profile of ST-11 for M16917. ST-11 is commonly associated with serogroup C in the United States (Harrison *et al.*, 2010). The CDC polyagglutination results did include a positive reaction with serogroup C antisera as well as reactivity with antisera from serogroups B, W135 and Y. However, the genome based MLST result is inconsistent with the SGS-PCR assay that resulted in a serogroup B result.

MLST is considered to be a reliable typing method in part because it utilizes seven loci spread across the *N. meningitidis* genome. However, the complete genome sequence should provide additional information with respect to the strain origins of M16917. Whole genome profiling of M16917 was performed using reference-based assembly against complete genome sequences representative of the four serogroups A, B, C and W135 (Tables 7.1 and 7.2). At the time of this analysis, there was no complete serogroup Y genome sequence. For each reference genome and for each corresponding reference nucleotide position, we ascertained the cumulative mapping quality score. Gaps were given a quality score of zero. Including gaps, the average quality score, *q,* was 56 ± 17 with a range of 0 to 64 where q=-10log(p), and where p is the probability of an incorrect base call. With this method, more distant genomes will have more areas not covered by the reads of query genome M16917. Fewer reads that map well to the reference genome will result in lower quality scores. Consequently, a closely related genome will have a relatively high overall quality score while a more distant genome will have a relatively low overall quality score. Whole genome profiling supports that M16917 is more closely related to the ST-11/serogroup C lineage, and is least related to the ST-74/serogroup B lineage (Figure 7.1 and Table 7.2). This result confirms the genome based MLST analysis and directly contradicts the SGS-PCR assay result.

ORIGIN OF THE M16917 CAPSULE POLYMERASE GENE

Taken together, the results of the MLST and the whole genome profiling analyses indicate that the genome of M16917 is most similar to serogroup C strains and most dissimilar to serogroup B. Thus, the serogroup B result from the SGS-PCR assay could be based on a false positive PCR, or more intriguingly, could point to a capsule switching event from serogroup C to B. Capsule switching occurs when a strain of *N. meningitidis* undergoes some kind of genetic alteration that results in the production of a capsule type that is more typical of a different serogroup. The genetic basis of capsule switching was first identified by Swartley et al. who characterized an isolate that underwent a switch from serogroup B to C based on a horizontal gene transfer and subsequent recombination event at the capsule locus (Swartley *et al.*, 1997). This particular transfer and recombination event introduced a serogroup C type capsule polymerase gene into a serogroup B genomic background. Subsequent studies have implied probable capsule switching events based on distinct genetic versus antigenic similarities of different serogroups (Harrison *et al.*, 2010; Tsang *et al.*, 2005; Vogel *et al.*, 2000).

The SGS-PCR assay is based on serogroup specific primer sets designed to the capsule polymerase gene found in the synthesis operon of the capsule locus (Mothershed *et al.*, 2004). To determine the origins of the M16917 capsule polymerase gene, we isolated its sequence from the genome assembly and compared it to reference capsule polymerase gene sequences from four different serogroups: A (strain Z2491), B (strain MC58), C (strain FAM18) and W135 (strain α275). Multiple sequence alignment and phylogenetic analysis of these sequences clearly indicates that the M16917 genome encodes a serogroup B type capsule polymerase gene (Figure 7.2). This result confirms the results of the SGS-PCR assay, and when considered together with the serogroup C origin of the M16917 genomic background, points to a possible C

to B capsule switching event. However, it is not immediately apparent why such a capsule switching event would produce a NG polyaggultination result in the SASG assay.

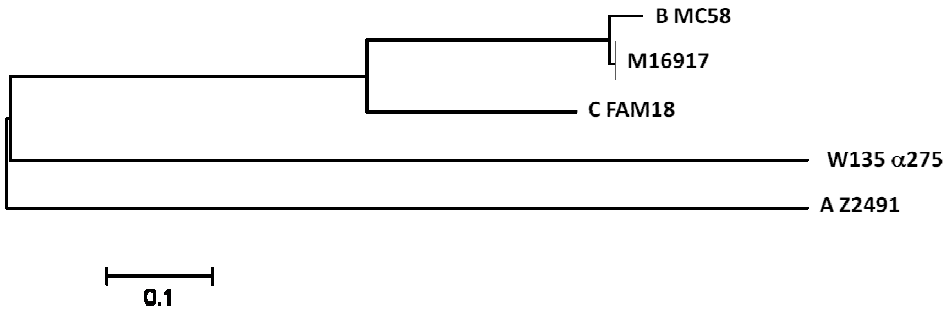


Figure 7.2. Phylogenetic analysis of the capsule polymerase gene. A phylogenetic tree was created with representative capsule polymerase genes of M16917 and those from serogroups A, B, C, and W135. The capsule polymerase gene of the nongroupable M16917 is most closely related to that of MC58 (serogroup B).

ORIGIN OF THE M16197 CAPSULE LOCUS

While a number of different cases of capsule switching have previously been identified, the nature of the recombination events that underlie these switches have not been precisely characterized (Beddek *et al.*, 2009; Swartley *et al.*, 1997; Tsang *et al.*, 2005). This is most likely due to the relatively limited amount of sequence information available in previous studies of capsule switching. Given our characterization of the complete genome sequence of M16917, along with the availability of complete genome sequences for serogroup B and C strains, we reasoned that we would have sufficient data to 1) evaluate whether capsule switching based on recombination had in fact occurred and 2) if so more precisely define the location of recombination breakpoints in the M16917 sequence. To do this, we constructed a three-way sequence alignment of the capsule loci from the M16917, MC58 (serogroup B) and FAM18 (serogroup C) genomes, and evaluated the distribution of sequence similarity between the M16917 sequence and the serogroup specific B or C sequences along the locus.

Alignment positions where M16917 agreed exclusively with either MC58 (B) or FAM18 (C) were considered to be informative with respect to the genomic origins of the sequence and were recorded into a string of informative positions (Figure 7.3A). A fifty-position sliding window was run along this string in single position steps, and the distribution of B and C informative positions in the upstream 25 positions versus the downstream 25 positions of each window was evaluated using 2 x 2 χ^2 contingency table (Figure 7.3B). Center window positions along the string that lead to the most disjoint B versus C distributions in the window, as measured by the significance level of the χ^2 test, are considered to represent most likely points of recombination. The χ^2 test Significance levels (*P*-values) were transformed into a negative-log-likelihood P-values to visually uncover local maxima. These local maxima represent the most likely recombination break points. Surprisingly, we uncovered four such sites that indicate the introduction of two serogroup B type genomic segments into the M16917 genome via recombination (Figure 7.3C). We had initially expected to observe that a single recombination event underlies the capsule switch as has been previously assumed (Beddek *et al.*, 2009; Swartley *et al.*, 1997; Tsang *et al.*, 2005). These results suggest that the process of recombination leading to capsule switching is even more dynamic than had been imagined. However, it is worth noting that the average size of a recombination even in *N. meningitidis* was previously inferred to be 1.1 kb (Jolley *et al.*, 2005), and the B serogroup specific sequence cassettes we observe are 1.9kb and 1.4kb in length consistent with this result. Moreover, from the recombinant breakpoint analysis, it can be seen that one of the inferred breakpoints is contained within the capsule polymerase gene (*synD* in Figure 7.3C). This would mean that the M16917 capsule polymerase gene, which is thought to be the primary determinant of the serogroup specific capsular structures, is likely to be a chimera containing both B and C type serogroup sequences.

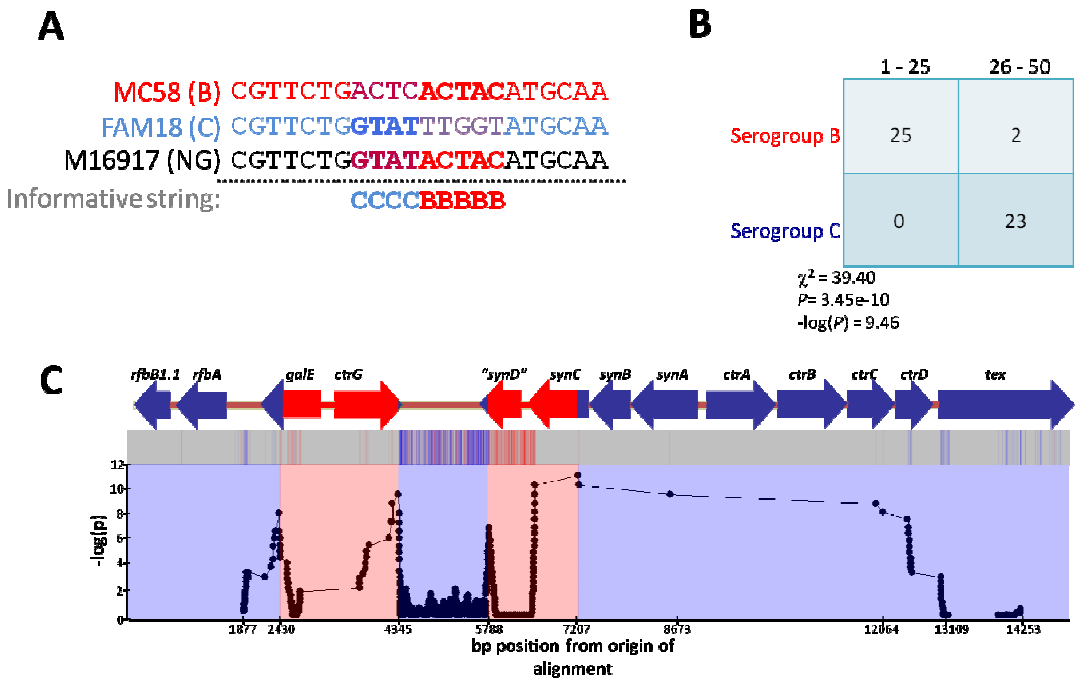


Figure 7.3. Capsule locus analysis. (A) Portion of the multiple sequence alignment of MC58, FAM18 and M16917 (serogroups B, C, and NG, respectively). The informative string is constructed by categorizing each informative site as either B or C depending on whether the nucleotide at that position on M16917 matched that of MC58 or FAM18. An informative site is formed by aligning M16917 capsule polymerase gene with reference capsule polymerase genes and recording the serogroup that M16917 agrees with (see Experimental Procedures). (B) At each site in the informative string, 25 sites to the left and right were observed and classified as B or C into a contingency 2x2 table. For statistical purposes, a pseudocount of 1 was added to each cell in the table (not shown). A $\chi 2$ analysis was performed, resulting in a probability value p. This was then transformed using –log(p). The contingency table displayed is for the local maxima shown at position 4345. (C) A graph of –log(p) vs nucleotide position, where p is derived from the contingency table as described in Fig. 7.3B. Directly above the graph are color-coded lines with a gray background that represent the informative string. At the top is the color-coded capsule locus as described by the informative string. Local maxima on the graph, derived from low p values, represent the most likely recombination points, which delineate regions in the genome that originate from either serogroup B (red) or serogroup C (blue). The mixture of serogroup B and C informative sites between *ctrG* and *siaD* could be an artifact in that it could be a result of having lower selective pressure in the intergenic regions.

Moreover, it is clear that the capsule polymerase gene of the isolate under investigation contains one such statistically significant breakpoint, making it a chimera.

The majority of the capsule polymerase gene, (*synD* in Figure 7.3C) and *synC* are included in the first cassette. SynD completes the synthesis of the polysaccharide capsule by polymerizing the monomeric sugar created by SynC (Edwards *et al.*, 1994; Warren & Blacklow, 1962). The second cassette includes the two genes *galE* and *ctrG*, formally known as NMB0065 (Hobb *et al.*, 2010). GalE sialylates lipooligosaccharide (LOS), and it has been shown that sialylation of LOS is indispensible for serum resistance for serogroup B *N. meningitidis* (Vogel *et al.*, 1997). CtrG functions with other proteins (CtrE, CtrF) as a chaperone in shuttling polysaccharide molecules to a transporter (comprised of CtrA-D), which then moves the polysaccharide to the surface to build the capsule.

CAPSULE SWITCHING AND NG POLYAGGULTINATION

Capsule switching has previously been implicated in the NG polyagglutination phenotype via changes to the capsule polymerase gene. Capsule polymerase genes complete the synthesis of polysaccharide capsules by polymerizing the monomeric sugar units from which they are composed in a serogroup specific fashion (Table 7.3). For example, the serogroup B type capsule polymerase SynD creates $\alpha 2 \rightarrow 8$ linkages between sialic acid monomers to form the B capsule polysaccharide, whereas the C type capsule polymerase gene encodes SynE that creates a capsule with $\alpha 2 \rightarrow 9$ sialic acid linkages (Frosch *et al.*, 1989; Swartley *et al.*, 1997; Swartley *et al.*, 1998). Steenbergen and Vimr experimentally demonstrated that the linkage specificity of identically composed *Escherichia coli* polysaccharide capsules could be changed from $\alpha 2 \rightarrow 8$ to $\alpha 2 \rightarrow 9$ by exchanging discrete regions of the polymerases that encode these specific linkages

(Steenbergen & Vimr, 2003). These results underscore the possibility that recombination events within capsule polymerase genes can lead to changes in capsule phenotpyes.

Table 7.3. Serogroup Capsule Type and Gene Target Names.

Serogroup	Capsule type	Capsule polymerase gene[a]	Ref
B	(α2→8)- N-acetylneuraminic acid	*synD*[b] *siaD* *siaD* of B $siaD_B$	(Frosch *et al.*, 1989; Swartley *et al.*, 1996)
C	(α2→9)- N-acetylneuraminic acid	*synE*[b] *siaD* of C $siaD_C$	(Frosch *et al.*, 1989; Swartley *et al.*, 1997)
W135	6-D-Gal(α1→4)-N-acetylneuraminic acid(α2→6)	*synG*[c] *siaD* of W135 $siaD_W$	(Bhattacharjee *et al.*, 1976; Claus *et al.*, 1997)
Y	6-D-Glc(α1→4)-N-acetylneuraminic acid(α2→6)	*synF*[c] *siaD* of Y $siaD_Y$	(Bhattacharjee *et al.*, 1976; Claus *et al.*, 1997)

[a]Bold names are primary names. Alternative names are given as unbolded.

[b]The *synD* and *synE* genes of serogroups B and C, respectively, are alleles and encode capsular polysaccharide polymerases that catalyze different linkages of sialic acid monomers (α2→8 linkage for serogroup B and α2→9 linkage for serogroup C).

[c]The *synF* and *synG* genes of serogroups Y and W135, respectively, are alleles and encode capsular polysaccharide polymerases that link heteropolymers of sialic acid to either glucose or galactose in serogroup Y or W135, respectively.

While interesting and potentially relevant, the aforementioned results were obtained in a laboratory setting on an analogous experimental system. In *N. meningitidis*, an actual isolate with a serogroup Y type genome was shown to be NG polyagglutinating by virtue of cross reactivity with both serogroup Y and W135 antisera (Tsang *et al.*, 2008). Amazingly, capsular structures produced by this strain were found to contain sialic acid units along with both glucose, which is a Y serogroup capsule specific sugar, and the W135 specific sugar and galactose. In other words, this strain was producing a hybrid capsule with both Y and W135

type features. The capsule polymerase gene was found to be clearly more closely related to the Y type gene than to the W135 capsule polymerase but differed from the canonical Y type sequence by three non-synonymous point mutations. These point mutations were apparently sufficient to produce a chimeric capsule and the resulting NG polyagglutinating phenotype.

Considering the results of these two studies together with the results we report here, it may be possible that recombination leading to a chimeric capsule polymerase gene could also lead to a hybrid capsule and the kind of NG polyagglutination result seen for M19617. Indeed, we found one of the inferred recombination breakpoints to be located within the capsule polymerase gene leading to a chimeric B/C polymerase (Figure 7.3C). It is tempting to speculate that such a chimeric polymerase gene may encode proteins with hybrid $\alpha 2 \rightarrow 8$ or $\alpha 2 \rightarrow 9$ sialic acid linkage activity leading capsular structures with cross reactivity to both B and C antisera. Confirmation of this hypothesis would require experimental characterization of the capsule structure of the M16917 isolate.

ACKNOWLEDGEMENTS

We thank the Illinois Department of Public Health, Chicago, Bacteriology Laboratory for donating isolate M16917 (also known as *N. meningitidis* 703526). This publication made use of the *Neisseria* Multi Locus Sequence Typing website (http://pubmlst.org/neisseria/) developed by Keith Jolley and Man-Suen Chan and sited at the University of Oxford. We thank Nancy Messonnier for reading and providing valuable feedback on this manuscript.

This work was supported by the Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology [BR-4839 to I.K.J.]; Georgia Research Alliance [GRA.VAC09.O to I.K.J., B.H.H., J.D.T., L.W.M., and L.S.K.]; and Bioinformatics program, Georgia Institute of Technology [to N.V.S.].

CHAPTER 8

CONCLUSIONS

This dissertation holds several chapters which have advanced the study of comparative genomics and epidemiology. Chapter 2 describes the online suite of tools called MGIP that accompany MLST analysis. Before MGIP, analysis of MLST data was very difficult and error-prone. Currently MGIP is being updated as indicated in chapter 3. One important update is to be able to accommodate other organisms including influenza and its sequence typing method STaRS. The interface for influenza is called InGen, and it has been showcased to U.S. labs that work with CDC and other health agencies. In addition, InGen has been presented to Egypt for a possible way to survey influenza and it has also been presented in a meeting in China where it may be accepted by international reference labs. InGen has a major advantage over other typing systems in that it will be used to gather influenza data in real time from participating typing institutions, thereby allowing CDC and other worldwide institutions to track influenza more quickly than ever.

Chapters 4 and 5 discuss the CG-Pipeline and NBase. These tools greatly facilitate the assembly, annotation, storage, visualization, and analyses of prokaryotic genomes. These tools will be taken to the next level, for they are continuously being updated. CG-Pipeline's updates will be to include different kinds of sequence data; to increase its prediction capabilities and accuracy; to increase the level of annotations; and to provide some genome-wide statistics. NBase will be updated to accommodate the amounts of data that will be generated by CG-Pipeline, which will be increased as whole-genome sequencing becomes more common and as the amount of data per genome will be increased. The analytic capability on NBase will be

increased. There are many tools for a GBrowse-based system including SynView that would be quite useful in our comparative genomics database (Wang *et al.*, 2006). In essence, these tools CG-Pipeline and NBase are quite useful as they are now, but we have plans to make them as comprehensive as possible.

In chapter 6, I asked a pressing question in meningococcal research: "What is the genomic basis of virulence in *N. meningitidis*?" Although some isolates cause devastating disease, the great majority do not cause disease and live as harmless commensals in humans. Scientists have wondered for almost a century what causes virulence in *N. meningitidis* and have been asking this question as it relates to its genome for the last ten years. I compared several whole virulent and carried genomes of *N. meningitidis* to find several discriminating SNPs and associated genes that serve as markers for virulence or carriage. These genes will be under further investigation of whether or not they can trigger meningococcal virulence. I have come closer to answering this question than anyone ever has.

In chapter 7, I asked another pressing question: "What causes nongroupability in *N. meningitidis*?" Determining the serogroup of an isolate is important, as it aids in surveillance and in determining vaccination strategy. When an isolate cannot be classified into any one group, it is considered nongroupable and hinders efforts to combat *N. meningitidis*. I compared the genome and capsule locus of a nongroupable isolate against several reference genomes and found homologous recombination events that probably led to its nongroupability. In addition, I created some algorithms to help in this investigation. The first, which I call genome profiling, shows degrees of relatedness when comparing a query genome against other reference genomes. The second shows exact recombination sites. These sites had never been found before this study despite other investigations into homologous recombination events in the capsule locus.

This thesis reflects my point of view of bioinformatics as it applies to public health and *Neisseria meningitidis*. I have developed important tools to fight this bacterium and have performed crucial analyses to aid in public health. These tools and studies will be able to be used to reduce bacterial meningitis worldwide and may be used to fight other infectious agents.

APPENDIX A

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

Table A.1. Loci that MGIP can analyze by default.

Locus Name[a]	Description[b]	Length (nts)[c]
Standard MLST loci		
abcZ	putative ABC transporter	433
adk_	adenylate kinase	465
aroE	shikimate dehydrogenase	490
fumC	fumarate hydratase	465
gdh_	glucose-6-phosphate dehydrogenase	501
pdhC	pyruvate dehydrogenase subunit	480
pgm_	phosphoglucomutase	450
Additional MGIP-specific loci		
fhb_	factor-H binding protein	*
gna2132	unknown	*
nadA	putative oligomeric coiled-coil adhesin	*
opa	opacity porin protein	*
porA	porin	*
porB	porin	*
fetA	iron-regulated outer membrane protein	*
penA	penicillin-binding protein	402

[a] Loci are given by their gene names.

[b] Names/descriptions of the encoded proteins.

[c] Fixed allele sequence lengths are shown; variable lengths are indicated by an asterisk (*).

Table A.2. Upload speeds for a set of trace files. The average size of a set of trace files in our tests was 11 megabytes (MB). These times do not include processing time.

Size (KB)	Speed (KB/second)	Label	Time (seconds)
11,000	10^6	Internal Network	0.011
11,000	10^4	T1	1.1
11,000	2000	FIOS	5.5
11,000	600	Cable	18.3
11,000	400	DSL	27.5
11,000	33	Dial-up	333

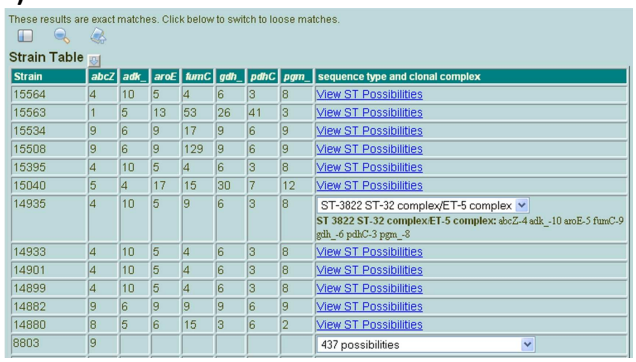
A)

These results are exact matches. Click below to switch to loose matches.

Strain Table

Strain	abcZ	adk_	aroE	fumC	gdh_	pdhC	pgm_	sequence type and clonal complex
15564	4	10	5	4	6	3	8	View ST Possibilities
15563	1	5	13	53	26	41	3	View ST Possibilities
15534	9	6	9	17	9	6	9	View ST Possibilities
15508	9	6	9	129	9	6	9	View ST Possibilities
15395	4	10	5	4	6	3	8	View ST Possibilities
15040	5	4	17	15	30	7	12	View ST Possibilities
14935	4	10	5	9	6	3	8	View ST Possibilities
14933	4	10	5	4	6	3	8	View ST Possibilities
14901	4	10	5	4	6	3	8	View ST Possibilities
14899	4	10	5	4	6	3	8	View ST Possibilities
14882	9	6	9	9	9	6	9	View ST Possibilities
14880	8	5	6	15	3	6	2	View ST Possibilities
8803	9							View ST Possibilities
4697			4					View ST Possibilities
4268			4					View ST Possibilities

B)

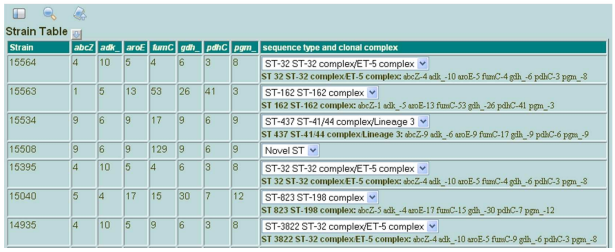


These results are exact matches. Click below to switch to loose matches.

Strain Table

Strain	abcZ	adk_	aroE	fumC	gdh_	pdhC	pgm_	sequence type and clonal complex
15564	4	10	5	4	6	3	8	View ST Possibilities
15563	1	5	13	53	26	41	3	View ST Possibilities
15534	9	6	9	17	9	6	9	View ST Possibilities
15508	9	6	9	129	9	6	9	View ST Possibilities
15395	4	10	5	4	6	3	8	View ST Possibilities
15040	5	4	17	15	30	7	12	View ST Possibilities
14935	4	10	5	9	6	3	8	ST-3822 ST-32 complex/ET-5 complex **ST 3822 ST-32 complex/ET-5 complex:** abcZ-4 adk_-10 aroE-5 fumC-9 gdh_-6 pdhC-3 pgm_-8
14933	4	10	5	4	6	3	8	View ST Possibilities
14901	4	10	5	4	6	3	8	View ST Possibilities
14899	4	10	5	4	6	3	8	View ST Possibilities
14882	9	6	9	9	9	6	9	View ST Possibilities
14880	8	5	6	15	3	6	2	View ST Possibilities
8803	9							437 possibilities

C)



Strain Table

Strain	abcZ	adk_	aroE	fumC	gdh_	pdhC	pgm_	sequence type and clonal complex
15564	4	10	5	4	6	3	8	ST-32 ST-32 complex/ET-5 complex **ST 32 ST-32 complex/ET-5 complex:** abcZ-4 adk_-10 aroE-5 fumC-4 gdh_-6 pdhC-3 pgm_-8
15563	1	5	13	53	26	41	3	ST-162 ST-162 complex **ST 162 ST-162 complex:** abcZ-1 adk_-5 aroE-13 fumC-53 gdh_-26 pdhC-41 pgm_-3
15534	9	6	9	17	9	6	9	ST-437 ST-41/44 complex/Lineage 3 **ST 437 ST-41/44 complex/Lineage 3:** abcZ-9 adk_-6 aroE-9 fumC-17 gdh_-9 pdhC-6 pgm_-9
15508	9	6	9	129	9	6	9	Novel ST
15395	4	10	5	4	6	3	8	ST-32 ST-32 complex/ET-5 complex **ST 32 ST-32 complex/ET-5 complex:** abcZ-4 adk_-10 aroE-5 fumC-4 gdh_-6 pdhC-3 pgm_-8
15040	5	4	17	15	30	7	12	ST-823 ST-198 complex **ST 823 ST-198 complex:** abcZ-5 adk_-4 aroE-17 fumC-15 gdh_-30 pdhC-7 pgm_-12
14935	4	10	5	9	6	3	8	ST-3822 ST-32 complex/ET-5 complex **ST 3822 ST-32 complex/ET-5 complex:** abcZ-4 adk_-10 aroE-5 fumC-9 gdh_-6 pdhC-3 pgm_-8

Figure A.1. The Strain Table. The strain table shows the MLST profile for each strain or isolate that has been analyzed by an individual user. (A) The strain table loads quickly because ST possibilities are not shown until the user makes a request by clicking on the "View ST Possibilities" link. (B) When there are multiple possibilities, all are shown in a drop down menu. When one of the possibilities is chosen, another request is made to the server to retrieve the ST and its allele components. Alternatively, an icon above the table allows the user to load every strain's ST possibilities at once. (C) One last notable item about the strain table is that it indicates novel STs when it encounters them (*e.g.* strain 15508).

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 4

Optional parts of the assembly stage included manual gap joining curation for scaffolding in the absence of paired-end reads, and frameshift detection for homopolymer-induced frameshifts.

The manual gap joining stage involved the layout of contigs according to their aligned position on the reference using the AMOS package and manual examination of each gap, adjacent contig alignments and reference annotation in the MAUVE visualization tool. We then recorded all gaps considered safe to join on the basis of this information into a gap fill specification file, which is a tabulated file in the format *"contig 1 name, contig 1 end position, reference start position, gap length, reference end position, contig 2 name, contig 2 start position"*, with one gap fill description per line. A script was then used to produce the final FASTA formatted output, with gaps filled with N (unknown nucleotides) by default, or optionally with sequence from the reference strain.

The homopolymer-induced frameshift stage used the FSFind package from (Kislyuk et al., 2009). Briefly, this package creates a GeneMark model of the genome, makes gene predictions, and then scans the genome for possible frameshift positions on the basis of ORF configuration and coding potential. Once the possible frameshift sites are identified, a putative translation of the protein possibly encoded by the broken gene is compared against a protein database (SwissProt by default). The predicted frameshift site is also scanned for adjacent homopolymers. A heuristic set of confidence score cutoffs is then used to provide a set of frameshift predictions while minimizing the false positive rate. The resulting homopolymer error predictions can be

used for either targeted re-sequencing or predictive correction using a supplied script. The output can be manually run through the gene prediction and annotation stages of the pipeline again.

To demonstrate the overall accuracy of the prediction stage, we ran it on the genome of *E. coli* K12, one of the best-annotated bacterial genomes. Our stage was able to detect 95.7% of intact ORFs annotated as protein-coding, and exactly predict starts in 85.5% of those. 50% of the predictions that do report incorrect start codons report starts within 35 nt of the true start, and all reported starts are within 200 nt of the true start.

The complete genome of *Escherichia coli* K12, accession number NC_000913.2, was downloaded from GenBank and its DNA sequence extracted into a FASTA file. The file was then given as input to the prediction component of the pipeline, which utilized the combination of *de novo* predictors GeneMark and Glimmer3. We note that the validation on the *E. coli* genome was performed without the BLAST gene prediction component, because all accurately annotated *E. coli* genes and genes of close relatives are present in the SwissProt database, so using that component would have made the trial biased. This is why the analysis proceeded with only the two *de novo* predictors, which impacted accuracy.

GenBank-formatted output of the prediction component was tabulated to include only CDS sequence annotation boundaries. The same procedure was done for the reference *E. coli* annotation from the original file. Sequences with frameshifted and interrupted CDS (*i.e.* non-intact ORFs) were omitted from the comparison due to lack of capability in our prediction component to detect such structures at this time.

APPENDIX C

SUPPLEMENTARY INFORMATION FOR CHAPTER 5

Table C.1. All genomes shown are included in the multiple sequence alignment and are present in NBase.

ID	Geographic Origin [a]	Year isolated [a]	Sg [b]	Disease	ST [c]	CC [c]	PMID [d]
M20918	Iowa, USA	2009	A	Invasive	4789	5	N/A
M13220	Philippines	2005	A	Invasive	7	5	20519285
M18575	Burkina Faso	2003	A	Invasive	2859	5	20519285
Z2491	Gambia	1983	A	Invasive	4	4	10761919
M17277	Maryland, USA	2006	NG	Carriage	5916	41/44	N/A
M16207	North Dakota, USA	2007	B	Invasive	162	162	N/A
M17094	Minnesota, USA	2008	B (C)	Carriage	32	32	N/A
MC58	Gloucester, UK	1985	B	Invasive	74	32	10710307
M10699	Oregon, USA	2003	B	Invasive	32	32	20519285
M5178	Oregon, USA	1998	B	Invasive	32	32	20519285
053442	Anhui province, China	2003	C	Invasive	4821	4821	18031983
8013	France	1989	C	Invasive	177	18	19818133
FAM18	North Carolina, USA	1980s	C	Invasive	11	11	17305430
M15141	New York, USA	2006	C	Invasive	11	11	20519285
α14	Bavaria, Germany	1999-2000	NG-*cnl*	Carriage	53	53	18305155
M17062	Minnesota, USA	2008	NG	Carriage	198	198	N/A
M15293	Georgia, USA	2006	NG (B)	Invasive	32	32	20519285
M16917	Illinois, USA	2007	NG	Invasive	11	11	N/A
α153	Bavaria, Germany	1999-2000	29E	Carriage	60	60	18305155
α275	Bavaria, Germany	1999-2000	W135	Carriage	22	22	18305155
M13519	New York, USA	2005	W135 (C)	Invasive	11	11	N/A
M17661	Michigan, USA	2008	W135 (C)	Invasive	11	11	N/A
M18774	Florida, USA	2009	W135	Invasive	11	11	N/A
NM9261	Burkina Faso	2002	W135	Invasive	11	11	20519285
M11791	New York, USA	2003	Y	Invasive	23	23	N/A
M14900	Oregon, USA	2006	Y	Invasive	1625	23	N/A
M20899	California, USA	2009	Y	Invasive	1624	167	N/A

Table C.1 Continued

[a] Origins are based on literature searches for each genome or based on unpublished notes at Centers for Disease Control and Prevention.

[b] Serogroup. Sgs are defined as a result of the SASG test, and if PCR had a differing result its resulting serogroup is in parentheses (Mothershed *et al.*, 2004). *cnl*: the capsule locus is nonexistent and no capsule is expressed.

[c] Sequence Type (ST), Clonal Complex (CC).

[d] PubMed ID of the genome announcement. N/A: not applicable; announced in this publication.

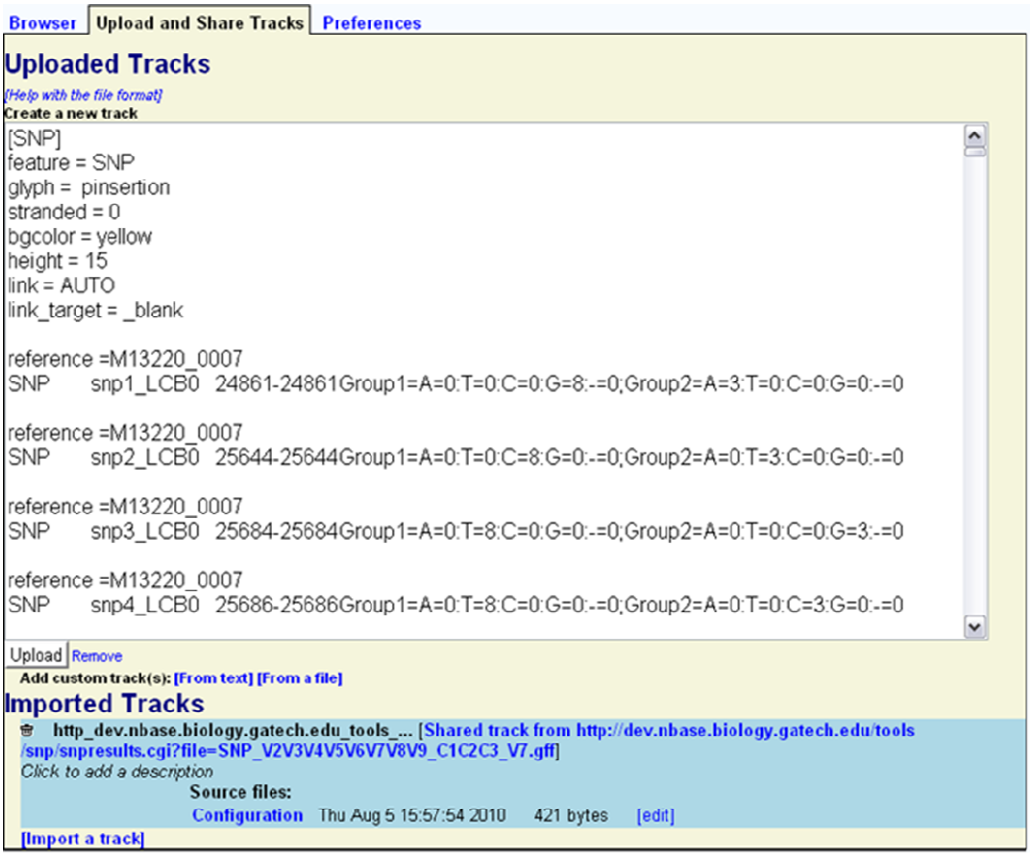


Figure C.1. Uploading additional tracks. The upload form, with the correct file format. The form provides instructions with a description of the file format.

APPENDIX D

SUPPLEMENTARY INFORMATION FOR CHAPTER 6

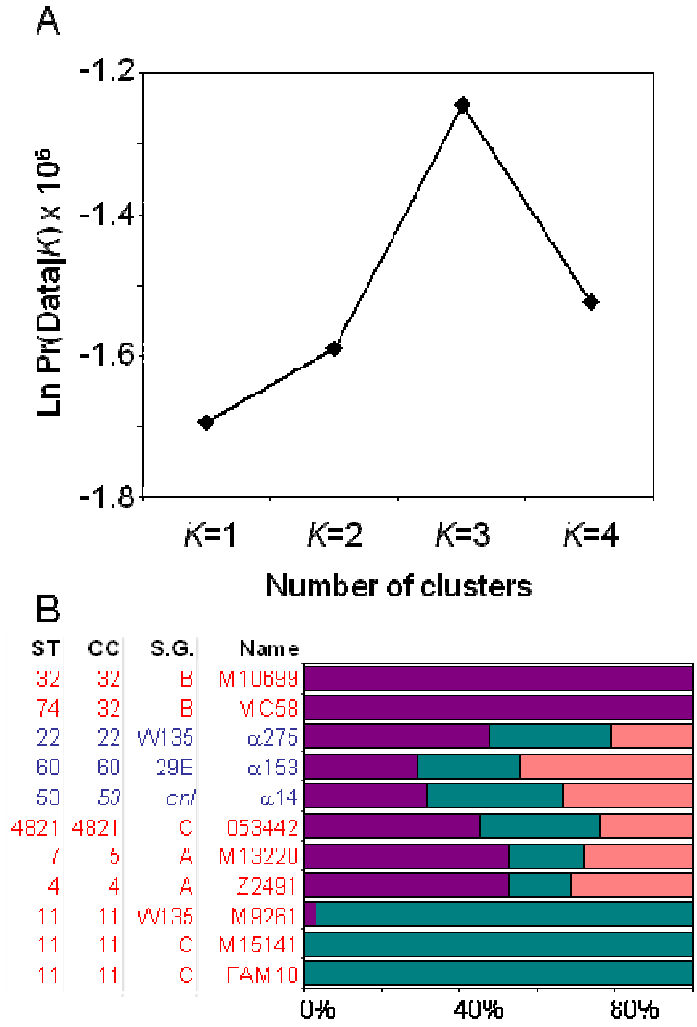


Figure D.1. Partitioning of SNP variation among *N. meningitidis* genomes using K-means clustering. (A) The likelihood of the observed SNP data based on the number of clusters *K*. (B) At *K*=3, the percentage of SNPs belonging to each cluster (purple, teal and peach) are shown for invasive (red) and carried (blue) strains. Each genome's ST, clonal complex, and serogroup is labeled.

File D.1. All SNP genes. All 113 SNP genes were characterized with descriptions and names of homologs from three reference genomes.

PUBLICATIONS

Katz, L. S., Bolen, C. R., Harcourt, B. H., Schmink, S., Wang, X., Kislyuk, A., Taylor, R. T., Mayer, L. W. & Jordan, I. K. (2009). Meningococcus genome informatics platform: a system for analyzing multilocus sequence typing data. *Nucleic Acids Res* 37, W606-611.

Katz, L. S., Humphrey, J. C., Conley, A. B. & other authors (2010). *Neisseria* Base: a comparative genomics database for *Neisseria meningitidis*. In preparation. *Database*.

Katz, L. S., Humphrey, J. C., Mayer, L. W. & Jordan, I. K. (2010). Update for CG-Pipeline: a computational pipeline for genome sequencing projects. In preparation.

Katz, L. S., Sharma, N. V., Harcourt, B. H., Thomas, J. D., Wang, X., Mayer, L. W. & Jordan, I. K. (2010). Using SNPs to Discriminate Disease Associated from Carried Genomes of *Neisseria meningitidis*. Submitted. *Journal of Bacteriology*.

Sharma, N.V., Katz, L. S., Rowe, L., Frace, M., Thomas, J. D., Harcourt, B. H., Mayer, L. W. & Jordan, I. K. (2010). The Genomic basis of a nongroupable *Neisseria meningitidis* isolate. In preparation.

Kislyuk, A. O., Katz, L. S., Agrawal, S. & other authors (2010). A computational genomics pipeline for prokaryotic sequencing projects. *Bioinformatics* **26**, 1819-1826.

REFERENCES

(2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* **38**, D142-148.

Achard, M. E., Hamilton, A. J., Dankowski, T., Heras, B., Schembri, M. S., Edwards, J. L., Jennings, M. P. & McEwan, A. G. (2009). A Periplasmic Thioredoxin-Like Protein Plays a Role in Defense against Oxidative Stress in *Neisseria gonorrhoeae*. *Infect Immun* **77**, 4934-4939.

Achtman, M. & Morelli, G. (2001). Pulsed-Field Gel Electrophoresis. In *Meningococcal Disease Methods and Protocols*, pp. 147-155. Edited by A. J. Pollard & M. C. J. Maiden. Totowa, New Jersey: Humana Press.

Aldea, M., Hernandez-Chico, C., de la Campa, A. G., Kushner, S. R. & Vicente, M. (1988). Identification, cloning, and expression of *bolA*, an *ftsZ*-dependent morphogene of *Escherichia coli*. *J Bacteriol* **170**, 5169-5176.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402.

Angellotti, M. C., Bhuiyan, S. B., Chen, G. & Wan, X. F. (2007). CodonO: codon usage bias analysis within and across genomes. *Nucleic Acids Res* **35**, W132-136.

Apweiler, R., Martin, M. J., O'Donovan, C., Magrane, M., Alam-Faruque, Y., Antunes, R. & *al., e.* (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* **38**, D142-148.

Arreaza, L. & Vázquez, J. A. (2001). Molecular Approach for the Study of Penicillin Resistance in *Neisseria meningitidis*. In *Meningococcal Disease Methods and Protocols*, pp. 107-119. Edited by A. J. Pollard & M. C. J. Maiden. Totowa, New Jersey: Humana Press.

Ashburner, M., Ball, C. A., Blake, J. A. & other authors (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29.

Ashton, F. E., Ryan, A., Diena, B. & Jennings, H. J. (1983). A new serogroup (L) of *Neisseria meningitidis*. *J Clin Microbiol* **17**, 722-727.

Aurrecoechea, C., Heiges, M., Wang, H. & other authors (2007). ApiDB: integrated resources for the apicomplexan bioinformatics resource center. *Nucleic Acids Res* **35**, D427-430.

Aziz, R., Bartels, D., Best, A. & other authors (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75.

Beddek, A. J., Li, M. S., Kroll, J. S., Jordan, T. W. & Martin, D. R. (2009). Evidence for capsule switching between carried and disease-causing *Neisseria meningitidis* strains. *Infect Immun* **77**, 2989-2994.

Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**, 783-795.

Benjamin, W. H., Jr., Yother, J., Hall, P. & Briles, D. E. (1991). The *Salmonella typhimurium* locus *mviA* regulates virulence in *Itys* but not *Ityr* mice: functional *mviA* results in avirulence; mutant (nonfunctional) *mviA* results in virulence. *J Exp Med* **174**, 1073-1083.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. (2010). GenBank. *Nucleic Acids Res* **38**, D46-51.

Bentley, D., Balasubramanian, S., Swerdlow, H. & other authors (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59.

Bentley, S. D., Vernikos, G. S., Snyder, L. A. & other authors (2007). Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. *PLoS Genet* **3**, e23.

Besemer, J., Lomsadze, A. & Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* **29**, 2607-2618.

Beucher, M. & Sparling, P. F. (1995). Cloning, sequencing, and characterization of the gene encoding FrpB, a major iron-regulated, outer membrane protein of *Neisseria gonorrhoeae*. *J Bacteriol* **177**, 2041-2049.

Bhattacharjee, A. K., Jennings, H. J., Kenny, C. P., Martin, A. & Smith, I. C. (1976). Structural determination of the polysaccharide antigens of *Neisseria meningitidis* serogroups Y, W-135, and BO1. *Can J Biochem* **54**, 1-8.

Bieri, T., Blasiar, D., Ozersky, P. & other authors (2007). WormBase: new content and better access. *Nucleic Acids Res* **35**, D506-510.

Bille, E., Zahar, J. R., Perrin, A. & other authors (2005). A chromosomally integrated bacteriophage in invasive meningococci. *J Exp Med* **201**, 1905-1913.

Bilukha, O. O. & Rosenstein, N. (2005). Prevention and control of meningococcal disease. Recommendations of the Advisory Committee on Immunization Practices (ACIP). *MMWR Recomm Rep* **54**, 1-21.

Boddicker, J. D. & Jones, B. D. (2004). Lon protease activity causes down-regulation of *Salmonella* pathogenicity island 1 invasion gene expression after infection of epithelial cells. *Infect Immun* **72**, 2002-2013.

Boeckmann, B., Bairoch, A., Apweiler, R. & other authors (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365-370.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421.

Campagne, G., Garba, A., Fabre, P. & other authors (2000). Safety and immunogenicity of three doses of a *Neisseria meningitidis* A + C diphtheria conjugate vaccine in infants from Niger. *Pediatr Infect Dis J* **19**, 144-150.

Capecchi, B., Adu-Bobie, J., Di Marcello, F., Ciucchi, L., Masignani, V., Taddei, A., Rappuoli, R., Pizza, M. & Arico, B. (2005). *Neisseria meningitidis* NadA is a new invasin which promotes bacterial adhesion to and penetration into human epithelial cells. *Mol Microbiol* **55**, 687-698.

Cartwright, K. (2006). Historical Aspects. In *Handbook of Meningococcal Disease Infection Biology, Vaccination, Clinical Management*, pp. 1-13. Edited by M. Frosch & M. C. J. Maiden. Weinheim, Germany: Wiley-VCH verlag GmbH & Co.

Castillo-Davis, C. I. & Hartl, D. L. (2003). GeneMerge--post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics* **19**, 891-892.

Caugant, D. A., Bovre, K., Gaustad, P., Bryn, K., Holten, E., Hoiby, E. A. & Froholm, L. O. (1986). Multilocus genotypes determined by enzyme electrophoresis of *Neisseria meningitidis* isolated from patients with systemic disease and from healthy carriers. *J Gen Microbiol* **132**, 641-652.

Chan, M. S. & Ventress, N. (2001).S.T.A.R.S. - Sequence Typing Analysis and Retrieval System.

Chen, I. & Dubnau, D. (2004). DNA uptake during bacterial transformation. *Nat Rev Microbiol* **2**, 241-249.

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y. & Jin, Q. (2005). VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res* **33**, D325-328.

Clamp, M., Cuff, J., Searle, S. M. & Barton, G. J. (2004). The Jalview Java alignment editor. *Bioinformatics* **20**, 426-427.

Claus, H., Vogel, U., Muhlenhoff, M., Gerardy-Schahn, R. & Frosch, M. (1997). Molecular divergence of the sia locus in different serogroups of *Neisseria meningitidis* expressing polysialic acid capsules. *Mol Gen Genet* **257**, 28-34.

Claus, H., Maiden, M. C., Maag, R., Frosch, M. & Vogel, U. (2002). Many carried meningococci lack the genes required for capsule synthesis and transport. *Microbiology* **148**, 1813-1819.

Claus, H., Maiden, M. C., Wilson, D. J., McCarthy, N. D., Jolley, K. A., Urwin, R., Hessler, F., Frosch, M. & Vogel, U. (2005). Genetic analysis of meningococci carried by children and young adults. *J Infect Dis* **191**, 1263-1271.

Cole, J. R., Wang, Q., Cardenas, E. & other authors (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37**, D141-145.

Conlin, C. A. & Miller, C. G. (2000). *opdA*, a *Salmonella enterica* serovar Typhimurium gene encoding a protease, is part of an operon regulated by heat shock. *J Bacteriol* **182**, 518-521.

Costerton, J. W., Irvin, R. T. & Cheng, K. J. (1981). The bacterial glycocalyx in nature and disease. *Annu Rev Microbiol* **35**, 299-324.

Danielson, L. & Mann, E. (1806). The history of a singlular and very mortal disease, which lately made its appearance in Medfield. *Med agric Reg* **1**, 65-69.

Darling, A. C., Mau, B., Blattner, F. R. & Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**, 1394-1403.

Davila, S., Wright, V. J., Khor, C. C. & other authors Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. *Nat Genet* **42**, 772–776.

de Souza, A. L. & Seguro, A. C. (2008). Two centuries of meningococcal infection: from Vieusseux to the cellular and molecular basis of disease. *J Med Microbiol* **57**, 1313-1321.

Dehal, P. S., Joachimiak, M. P., Price, M. N. & other authors (2010). MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res* **38**, D396-400.

Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**, 4636-4641.

Didelot, X., Urwin, R., Maiden, M. C. & Falush, D. (2009). Genealogical typing of *Neisseria meningitidis*. *Microbiology* **155**, 3176-3186.

Dolan-Livengood, J. M., Miller, Y. K., Martin, L. E., Urwin, R. & Stephens, D. S. (2003). Genetic basis for nongroupable *Neisseria meningitidis*. *J Infect Dis* **187**, 1616-1628.

Drysdale, R. (2008). FlyBase : a database for the *Drosophila* research community. *Methods Mol Biol* **420**, 45-59.

Du, Y., Lenz, J. & Arvidson, C. G. (2005). Global gene expression and the role of sigma factors in *Neisseria gonorrhoeae* in interactions with epithelial cells. *Infect Immun* **73**, 4834-4845.

Dull, P. M., Abdelwahab, J., Sacchi, C. T. & other authors (2005). *Neisseria meningitidis* serogroup W-135 carriage among US travelers to the 2001 Hajj. *J Infect Dis* **191**, 33-39.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797.

Edwards, U., Muller, A., Hammerschmidt, S., Gerardy-Schahn, R. & Frosch, M. (1994). Molecular analysis of the biosynthesis pathway of the alpha-2,8 polysialic acid capsule by *Neisseria meningitidis* serogroup B. *Mol Microbiol* **14**, 141-149.

Eid, J., Fehr, A., Gray, J. & other authors (2009). Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138.

Elsik, C. G., Worley, K. C., Zhang, L. & other authors (2006). Community annotation: procedures, protocols, and supporting tools. *Genome Res* **16**, 1329-1333.

Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* **2**, 953-971.

Enright, M. C. & Spratt, B. G. (1999). Multilocus sequence typing. *Trends Microbiol* **7**, 482-487.

Estabrook, M. M., Jack, D. L., Klein, N. J. & Jarvis, G. A. (2004). Mannose-binding lectin binds to two major outer membrane proteins, opacity protein and porin, of *Neisseria meningitidis*. *J Immunol* **172**, 3784-3792.

Ewing, B. & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 186-194.

Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**, 175-185.

Finne, J., Leinonen, M. & Makela, P. H. (1983). Antigenic similarities between brain components and bacteria causing meningitis. Implications for vaccine development and pathogenesis. *Lancet* **2**, 355-357.

Fleischmann, R. D., Adams, M. D., White, O. & other authors (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512.

Fletcher, L. D., Bernfield, L., Barniak, V. & other authors (2004). Vaccine potential of the *Neisseria meningitidis* 2086 lipoprotein. *Infect Immun* **72**, 2088-2100.

Flexner, S. (1913). The Results of the Serum Treatment in Thirteen Hundred Cases of Epidemic Meningitis. *J Exp Med* **17**, 553-576.

Flicek, P., Aken, B. L., Ballester, B. & other authors (2010). Ensembl's 10th year. *Nucleic Acids Res* **38**, D557-562.

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res* **32**, W273-279.

Frosch, M., Weisgerber, C. & Meyer, T. F. (1989). Molecular characterization and expression in *Escherichia coli* of the gene complex encoding the polysaccharide capsule of *Neisseria meningitidis* group B. *Proc Natl Acad Sci U S A* **86**, 1669-1673.

Frosch, M. & Vogel, U. (2006). *Structure and Genetics of the Meningococcal Capsule*. In *Handbook of Meningococcal Disease Infection Biology, Vaccination, Clinical Management*, pp. 145-162. Edited by M. Frosch & M. C. J. Maiden. Weinheim, Germany: Wiley-VCH verlag GmbH & Co.

Geoffroy, M. C., Floquet, S., Metais, A., Nassif, X. & Pelicic, V. (2003). Large-scale analysis of the meningococcus genome by gene disruption: resistance to complement-mediated lysis. *Genome Res* **13**, 391-398.

Gerlach, G., von Wintzingerode, F., Middendorf, B. & Gross, R. (2001). Evolutionary trends in the genus *Bordetella*. *Microbes Infect* **3**, 61-72.

Goldschneider, I., Gotschlich, E. C. & Artenstein, M. S. (1969a). Human immunity to the meningococcus. II. Development of natural immunity. *J Exp Med* **129**, 1327-1348.

Goldschneider, I., Gotschlich, E. C. & Artenstein, M. S. (1969b). Human immunity to the meningococcus. I. The role of humoral antibodies. *J Exp Med* **129**, 1307-1326.

Goodman, S. D. & Scocca, J. J. (1988). Identification and arrangement of the DNA sequence recognized in specific transformation of *Neisseria gonorrhoeae*. *Proc Natl Acad Sci U S A* **85**, 6982-6986.

Gordon, D., Desmarais, C. & Green, P. (2001). Automated finishing with autofinish. *Genome Res* **11**, 614-625.

Gotschlich, E. C., Goldschneider, I. & Artenstein, M. S. (1969a). Human immunity to the meningococcus. V. The effect of immunization with meningococcal group C polysaccharide on the carrier state. *J Exp Med* **129**, 1385-1395.

Gotschlich, E. C., Goldschneider, I. & Artenstein, M. S. (1969b). Human immunity to the meningococcus. IV. Immunogenicity of group A and group C meningococcal polysaccharides in human volunteers. *J Exp Med* **129**, 1367-1384.

Gotschlich, E. C., Liu, T. Y. & Artenstein, M. S. (1969c). Human immunity to the meningococcus. 3. Preparation and immunochemical properties of the group A, group B, and group C meningococcal polysaccharides. *J Exp Med* **129**, 1349-1365.

Harrison, L. H., Trotter, C. L. & Ramsay, M. E. (2009). Global epidemiology of meningococcal disease. *Vaccine* **27 Suppl 2**, B51-63.

Harrison, L. H., Shutt, K. A., Schmink, S. E. & other authors (2010). Population structure and capsular switching of invasive *Neisseria meningitidis* isolates in the pre-meningococcal conjugate vaccine era--United States, 2000-2005. *J Infect Dis* **201**, 1208-1224.

Hauck, C. R. & Meyer, T. F. (1997). The lysosomal/phagosomal membrane protein h-lamp-1 is a target of the IgA1 protease of *Neisseria gonorrhoeae*. *FEBS Lett* **405**, 86-90.

Hobb, R. I., Tzeng, Y. L., Choudhury, B. P., Carlson, R. W. & Stephens, D. S. (2010). Requirement of NMB0065 for connecting assembly and export of sialic acid capsular polysaccharides in *Neisseria meningitidis*. *Microbes Infect* **12**, 476-487.

Holmes, E. C., Urwin, R. & Maiden, M. C. (1999). The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. *Mol Biol Evol* **16**, 741-749.

Hotopp, J. C., Grifantini, R., Kumar, N. & other authors (2006). Comparative genomics of *Neisseria meningitidis*: core genome, islands of horizontal transfer and pathogen-specific genes. *Microbiology* **152**, 3733-3749.

Hudson, R. R., Slatkin, M. & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583-589.

Ingmer, H. & Brøndsted, L. (2009). Proteases in bacterial pathogenesis. *Res Microbiol* **160**, 704-710.

Inoue, A., Murata, Y., Takahashi, H., Tsuji, N., Fujisaki, S. & Kato, J. (2008). Involvement of an essential gene, *mviN*, in murein synthesis in *Escherichia coli*. *J Bacteriol* **190**, 7298-7301.

Jacobsson, S., Hedberg, S. T., Molling, P., Unemo, M., Comanducci, M., Rappuoli, R. & Olcen, P. (2009). Prevalence and sequence variations of the genes encoding the five antigens included in the novel 5CVMB vaccine covering group B meningococcal disease. *Vaccine* **27**, 1579-1584.

Jarva, H., Ram, S., Vogel, U., Blom, A. M. & Meri, S. (2005). Binding of the complement inhibitor C4bp to serogroup B *Neisseria meningitidis*. *J Immunol* **174**, 6299-6307.

Jolley, K. A. (2001). Multi-Locus Sequence Typing. In *Meningococcal Disease Methods and Protocols*, pp. 173-186. Edited by A. J. Pollard & M. C. J. Maiden. Totowa, New Jersey: Humana Press.

Jolley, K. A., Chan, M. S. & Maiden, M. C. (2004). mlstdbNet - distributed multi-locus sequence typing (MLST) databases. *BMC Bioinformatics* **5**, 86.

Jolley, K. A., Wilson, D. J., Kriz, P., McVean, G. & Maiden, M. C. (2005). The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. *Mol Biol Evol* **22**, 562-569.

Jolley, K. A., Gray, S. J., Suker, J. & Urwin, R. (2006). Methods for Typing of Meningococci. In *Handbook of Meningococcal Disease Infection Biology, Vaccination, Clinical Management*, pp. 37-51. Edited by M. Frosch & M. C. J. Maiden. Weinheim, Germany: Wiley-VCH verlag GmbH & Co.

Jolley, K. A., Brehony, C. & Maiden, M. C. (2007). Molecular typing of meningococci: recommendations for target choice and nomenclature. *FEMS Microbiol Rev* **31**, 89-96.

Joseph, B., Schneiker-Bekel, S., Schramm-Gluck, A. & other authors (2010). Comparative genome biology of a serogroup B carriage and disease strain supports a polygenic nature of meningococcal virulence. *J Bacteriol* **192**, 5363-5377.

Juncker, A. S., Willenbrock, H., Von Heijne, G., Brunak, S., Nielsen, H. & Krogh, A. (2003). Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci* **12**, 1652-1662.

Katz, L. S., Bolen, C. R., Harcourt, B. H., Schmink, S., Wang, X., Kislyuk, A., Taylor, R. T., Mayer, L. W. & Jordan, I. K. (2009). Meningococcus genome informatics platform: a system for analyzing multilocus sequence typing data. *Nucleic Acids Res* **37**, W606-611.

Katz, L. S., Humphrey, J. C., Conley, A. B. & other authors (2010). *Neisseria* Base: a comparative genomics database for *Neisseria meningitidis*. In submission. *Database*.

Kawai, M., Nakao, K., Uchiyama, I. & Kobayashi, I. (2006). How genomes rearrange: genome comparison within bacteria *Neisseria* suggests roles for mobile elements in formation of complex genome polymorphisms. *Gene* **383**, 52-63.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. & Haussler, D. (2002). The human genome browser at UCSC. *Genome Res* **12**, 996-1006.

Kislyuk, A., Lomsadze, A., Lapidus, A. & Borodovsky, M. (2009). Frameshift detection in prokaryotic genomic sequences. *International journal of bioinformatics research and applications* **5**, 458-477.

Kislyuk, A. O., Katz, L. S., Agrawal, S. & other authors (2010). A computational genomics pipeline for prokaryotic sequencing projects. *Bioinformatics* **26**, 1819-1826.

Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567-580.

Kroll, J. S., Wilks, K. E., Farrant, J. L. & Langford, P. R. (1998). Natural genetic exchange between *Haemophilus* and *Neisseria*: intergeneric transfer of chromosomal genes between major human pathogens. *Proc Natl Acad Sci U S A* **95**, 12381-12385.

Kumar, S., Nei, M., Dudley, J. & Tamura, K. (2008). MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief Bioinform* **9**, 299-306.

Kuo, A. & Grigoriev, I. (2009). Challenges in Whole-Genome Annotation of Pyrosequenced Fungal Genomes.

Lagesen, K., Hallin, P., Rodland, E. A., Staerfeldt, H. H., Rognes, T. & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35**, 3100-3108.

Lander, E. S. & Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231-239.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25.

Lapierre, P. & Gogarten, P. (2009). Estimating the size of the bacterial pan-genome. *Trends in Genetics* **25**, 107-110.

Larkin, M. A., Blackshields, G., Brown, N. P. & other authors (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948.

Lin, L., Ayala, P., Larson, J., Mulks, M., Fukuda, M., Carlsson, S. R., Enns, C. & So, M. (1997). The *Neisseria* type 2 IgA1 protease cleaves LAMP1 and promotes survival of bacteria within epithelial cells. *Mol Microbiol* **24**, 1083-1094.

Lingappa, J. R., Al-Rabeah, A. M., Hajjeh, R. & other authors (2003). Serogroup W-135 meningococcal disease during the Hajj, 2000. *Emerg Infect Dis* **9**, 665-671.

Linz, B., Schenker, M., Zhu, P. & Achtman, M. (2000). Frequent interspecific genetic exchange between commensal Neisseriae and *Neisseria meningitidis*. *Mol Microbiol* **36**, 1049-1058.

Litt, D. J., Savino, S., Beddek, A. & other authors (2004). Putative vaccine antigens from *Neisseria meningitidis* recognized by serum antibodies of young children convalescing after meningococcal disease. *J Infect Dis* **190**, 1488-1497.

Lowe, T. M. & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-964.

Lucidarme, J., Comanducci, M., Findlow, J. & other authors (2009). Characterization of *fHbp*, *nhba* (*gna2132*), *nadA*, *porA*, sequence type (ST), and genomic presence of IS1301 in group B meningococcal ST269 clonal complex isolates from England and Wales. *J Clin Microbiol* **47**, 3577-3585.

Lukashin, A. V. & Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* **26**, 1107-1115.

MacCallum, I., Przybylski, D., Gnerre, S. & other authors (2009). ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. *Genome Biology* **10**, R103.

Madico, G., Welsch, J. A., Lewis, L. A. & other authors (2006). The meningococcal vaccine candidate GNA1870 binds the complement regulatory protein factor H and enhances serum resistance. *J Immunol* **177**, 501-510.

Mahillon, J. & Chandler, M. (1998). Insertion sequences. *Microbiol Mol Biol Rev* **62**, 725-774.

Maiden, M. C., Bygraves, J. A., Feil, E. & other authors (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* **95**, 3140-3145.

Maiden, M. C. & Jolley, K. A. (2010).*Neisseria* population genomics: integrating whole genome data with multi locus approaches to epidemiology and population biology. In *17th International Pathogenic Neisseria Conference*, pp. 47. Banff, Alberta, Canada.

Maiden, M. C. J. & Caugant, D. A. (2006). The Population Biology of *Neisseria meningitidis*: Implications for Meningococcal Disease, Epidemiology and Control. In *Handbook of Meningococcal Disease Infection Biology, Vaccination, Clinical Management*, pp. 17-35. Edited by M. Frosch & M. C. J. Maiden. Weinheim, Germany: Wiley-VCH verlag GmbH & Co.

Marceau, M., Forest, K., Beretti, J. L., Tainer, J. & Nassif, X. (1998). Consequences of the loss of O-linked glycosylation of meningococcal type IV pilin on piliation and pilus-mediated adhesion. *Mol Microbiol* **27**, 705-715.

Margulies, M., Egholm, M., Altman, W. E. & other authors (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380.

Markowitz, V., Chen, I. M., Palaniappan, K. & other authors (2009). The integrated microbial genomes system: an expanding comparative analysis resource. *Nucl Acids Res*, gkp887.

Masignani, V., Comanducci, M., Giuliani, M. M. & other authors (2003). Vaccination against *Neisseria meningitidis* using three variants of the lipoprotein GNA1870. *J Exp Med* **197**, 789-799.

Mayer, L. W., Reeves, M. W., Al-Hamdan, N. & other authors (2002). Outbreak of W135 meningococcal disease in 2000: not emergence of a new W135 strain but clonal expansion within the electophoretic type-37 complex. *J Infect Dis* **185**, 1596-1605.

Mayor, C., Brudno, M., Schwartz, J. R., Poliakov, A., Rubin, E. M., Frazer, K. A., Pachter, L. S. & Dubchak, I. (2000). VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**, 1046-1047.

McGee, Z. A., Stephens, D. S., Hoffman, L. H., Schlech, W. F., 3rd & Horn, R. G. (1983). Mechanisms of mucosal invasion by pathogenic *Neisseria*. *Rev Infect Dis* **5 Suppl 4**, S708-714.

Meyers, L. A., Levin, B. R., Richardson, A. R. & Stojiljkovic, I. (2003). Epidemiology, hypermutation, within-host evolution and the virulence of *Neisseria meningitidis*. *Proc Biol Sci* **270**, 1667-1677.

Miller, E., Salisbury, D. & Ramsay, M. (2001). Planning, registration, and implementation of an immunisation campaign against meningococcal serogroup C disease in the UK: a success story. *Vaccine* **20 Suppl 1**, S58-67.

Miller, J., Delcher, A., Koren, S. & other authors (2008). Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818-2824.

Mistry, D. & Stockley, R. A. (2006). IgA1 protease. *Int J Biochem Cell Biol* **38**, 1244-1248.

Mothershed, E. A., Sacchi, C. T., Whitney, A. M. & other authors (2004). Use of real-time PCR to resolve slide agglutination discrepancies in serogroup identification of *Neisseria meningitidis*. *J Clin Microbiol* **42**, 320-328.

Mulder, N. & Apweiler, R. (2007). InterPro and InterProScan: Tools for Protein Sequence Classification and Comparison. *Methods Mol Biol* **396**, 59-70.

Murphy, E., Andrew, L., Lee, K. L. & other authors (2009). Sequence diversity of the factor H binding protein vaccine candidate in epidemiologically relevant strains of serogroup B *Neisseria meningitidis*. *J Infect Dis* **200**, 379-389.

Nassif, X., Marceau, M., Pujol, C., Pron, B., Beretti, J. L. & Taha, M. K. (1997). Type-4 pili and meningococcal adhesiveness. *Gene* **192**, 149-153.

Nei, M. & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**, 418-426.

Palmgren, H. (2009). Meningococcal disease and climate. *Glob Health Action* **2**.

Parkhill, J., Achtman, M., James, K. D. & other authors (2000). Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* **404**, 502-506.

Parkhill, J., Sebaihia, M., Preston, A. & other authors (2003). Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat Genet* **35**, 32-40.

Peng, J., Yang, L., Yang, F. & other authors (2008). Characterization of ST-4821 complex, a unique *Neisseria meningitidis* clone. *Genomics* **91**, 78-87.

Perrin, A., Bonacorsi, S., Carbonnelle, E., Talibi, D., Dessen, P., Nassif, X. & Tinsley, C. (2002). Comparative genomics identifies the genetic islands that distinguish *Neisseria meningitidis*, the agent of cerebrospinal meningitis, from other *Neisseria* species. *Infect Immun* **70**, 7063-7072.

Pizza, M., Scarlato, V., Masignani, V. & other authors (2000). Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* **287**, 1816-1820.

Platt, S., Pichon, B., George, R. & Green, J. (2006). A bioinformatics pipeline for high-throughput microbial multilocus sequence typing (MLST) analyses. *Clin Microbiol Infect* **12**, 1144-1146.

Pop, M., Phillippy, A., Delcher, A. L. & Salzberg, S. L. (2004). Comparative genome assembly. *Brief Bioinform* **5**, 237-248.

Popovic, T., Ajello, G. W. & Facklam, R. (1999). Laboratory Manual for the Diagnosis of Meningitis caused by *Neisseria meningitidis*, *Streptococcus pneumoniae*

and *Haemophilus influenzae*. *World Health Organization Communicable Disease Surveillance and Response*.

Popovic, T. & Ajello, G. W. (2003). *Neisseria meningitidis: Confirmatory Identification and Antimicrobial Susceptibility Testing*. Geneva: World Health Organization.

Power, P. M. & Jennings, M. P. (2003). The genetics of glycosylation in Gram-negative bacteria. *FEMS Microbiol Lett* **218**, 211-222.

Price, M. N., Huang, K. H., Alm, E. J. & Arkin, A. P. (2005). A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res* **33**, 880-892.

Pritchard, J. K., Stephens, M. & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Pruitt, K. D., Tatusova, T. & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-65.

Quinlan, A., Stewart, D., Stromberg, M. & Marth, G. (2008). Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat Meth* **5**, 179-181.

Racoosin, J. A., Whitney, C. G., Conover, C. S. & Diaz, P. S. (1998). Serogroup Y meningococcal disease in Chicago, 1991-1997. *JAMA* **280**, 2094-2098.

Rappuoli, R. (2000). Reverse vaccinology. *Curr Opin Microbiol* **3**, 445-450.

Rappuoli, R. (2001). Reverse vaccinology, a genome-based approach to vaccine development. *Vaccine* **19**, 2688-2691.

Rinaudo, C. D., Telford, J. L., Rappuoli, R. & Seib, K. L. (2009). Vaccinology in the genome era. *J Clin Invest* **119**, 2515-2525.

Rissman, A., Mau, B., Biehl, B., Darling, A., Glasner, J. & Perna, N. (2009). Reordering contigs of draft genomes using the Mauve Aligner. *Bioinformatics* **25**, 2071-2073.

Rosenstein, N. E., Perkins, B. A., Stephens, D. S., Popovic, T. & Hughes, J. M. (2001). Meningococcal disease. *N Engl J Med* **344**, 1378-1388.

Rusniok, C., Vallenet, D., Floquet, S. & other authors (2009). NeMeSys: a biological resource for narrowing the gap between sequence and function in the human pathogen *Neisseria meningitidis*. *Genome Biol* **10**, R110.

Sacchi, C. T., Lemos, A. P., Whitney, A. M., Solari, C. A., Brandt, M. E., Melles, C. E., Frasch, C. E. & Mayer, L. W. (1998). Correlation between serological and sequencing analyses of the PorB outer membrane protein in the *Neisseria meningitidis* serotyping system. *Clin Diagn Lab Immunol* **5**, 348-354.

Sacchi, C. T., Whitney, A. M., Popovic, T. & other authors (2000). Diversity and prevalence of PorA types in *Neisseria meningitidis* serogroup B in the United States, 1992-1998. *J Infect Dis* **182**, 1169-1176.

Sadarangani, M. & Pollard, A. J. (2010). Serogroup B meningococcal vaccines-an unfinished story. *Lancet Infect Dis* **10**, 112-124.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406-425.

Sampath, R., Russell, K. L., Massire, C. & other authors (2007). Global surveillance of emerging Influenza virus genotypes by mass spectrometry. *PLoS One* **2**, e489.

Santos, J. M., Lobo, M., Matos, A. P., De Pedro, M. A. & Arraiano, C. M. (2002). The gene *bolA* regulates *dacA* (PBP5), *dacC* (PBP6) and *ampC* (AmpC), promoting normal morphology in *Escherichia coli*. *Mol Microbiol* **45**, 1729-1740.

Sayers, E. W., Barrett, T., Benson, D. A. & other authors (2010). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **38**, D5-16.

Scarselli, M., Serruto, D., Montanari, P., Capecchi, B., Adu-Bobie, J., Veggi, D., Rappuoli, R., Pizza, M. & Arico, B. (2006). *Neisseria meningitidis* NhhA is a multifunctional trimeric autotransporter adhesin. *Mol Microbiol* **61**, 631-644.

Schneider, M. C., Exley, R. M., Chan, H., Feavers, I., Kang, Y. H., Sim, R. B. & Tang, C. M. (2006). Functional significance of factor H binding to *Neisseria meningitidis*. *J Immunol* **176**, 7566-7575.

Schoen, C. & Claus, H. (2006). *Neisseria meningitidis* Genome Sequencing Projects. In *Handbook of Meningococcal Disease Infection Biology, Vaccination, Clinical Management*, pp. 77-79. Edited by M. Frosch & M. C. J. Maiden. Weinheim, Germany: Wiley-VCH verlag GmbH & Co.

Schoen, C., Blom, J., Claus, H. & other authors (2008). Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in *Neisseria meningitidis*. *Proc Natl Acad Sci U S A* **105**, 3473-3478.

Seib, K. L., Oriente, F., Adu-Bobie, J., Montanari, P., Ferlicca, F., Giuliani, M. M., Rappuoli, R., Pizza, M. & Delany, I. (2010). Influence of serogroup B meningococcal vaccine antigens on growth and survival of the meningococcus in vitro and in ex vivo and in vivo models of infection. *Vaccine* **28**, 2416-2427.

Serruto, D., Spadafina, T., Ciucchi, L. & other authors (2010). *Neisseria meningitidis* GNA2132, a heparin-binding protein that induces protective immunity in humans. *Proc Natl Acad Sci U S A* **107**, 3770-3775.

Seshadri, R., Kravitz, S., Smarr, L., Gilna, P. & Frazier, M. (2007). CAMERA: A Community Resource for Metagenomics. *PLoS Biol* **5**, e75.

Sexton, K., Lennon, D., Oster, P., Aaberge, I., Martin, D., Reid, S., Wong, S. & O'Hallahan, J. (2004a). Proceedings of the Meningococcal Vaccine Strategy World Health Organization satellite meeting, 10 March 2004, Auckland, New Zealand. *N Z Med J* **117**, 1 p preceding U1027.

Sexton, K., Lennon, D., Oster, P. & other authors (2004b). The New Zealand Meningococcal Vaccine Strategy: a tailor-made vaccine to combat a devastating epidemic. *N Z Med J* **117**, U1015.

Shendure, J., Porreca, G. J., Reppas, N. B. & other authors (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-1732.

Sjolinder, H., Eriksson, J., Maudsdotter, L., Aro, H. & Jonsson, A. B. (2008). Meningococcal outer membrane protein NhhA is essential for colonization and disease by preventing phagocytosis and complement attack. *Infect Immun* **76**, 5412-5420.

Smith, H. O., Gwinn, M. L. & Salzberg, S. L. (1999). DNA uptake signal sequences in naturally transformable bacteria. *Res Microbiol* **150**, 603-616.

Snyder, L. A. & Saunders, N. J. (2006). The majority of genes in the pathogenic *Neisseria* species are present in non-pathogenic *Neisseria lactamica*, including those designated as 'virulence genes'. *BMC Genomics* **7**, 128.

Sommer, D. D., Delcher, A. L., Salzberg, S. L. & Pop, M. (2007). Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* **8**, 64.

Stabler, R. & Hinds, J. (2006). The majority of genes in the pathogenic *Neisseria* species are present in non-pathogenic *Neisseria lactamica*, including those designated as virulence genes: response. *BMC Genomics* **7**, 129.

Stabler, R. A., Marsden, G. L., Witney, A. A., Li, Y., Bentley, S. D., Tang, C. M. & Hinds, J. (2005). Identification of pathogen-specific genes through microarray analysis of pathogenic and commensal *Neisseria* species. *Microbiology* **151**, 2907-2922.

Stajich, J. E., Block, D., Boulez, K. & other authors (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12**, 1611-1618.

Steenbergen, S. M. & Vimr, E. R. (2003). Functional relationships of the sialyltransferases involved in expression of the polysialic acid capsules of *Escherichia coli* K1 and K92 and *Neisseria meningitidis* groups B or C. *J Biol Chem* **278**, 15349-15359.

Stein, L. D., Mungall, C., Shu, S. & other authors (2002). The generic genome browser: a building block for a model organism system database. *Genome Res* **12**, 1599-1610.

Stephens, D. S., Hoffman, L. H. & McGee, Z. A. (1983). Interaction of *Neisseria meningitidis* with human nasopharyngeal mucosa: attachment and entry into columnar epithelial cells. *J Infect Dis* **148**, 369-376.

Stewart, A., Osborne, B. & Read, T. (2009). DIYA: a bacterial annotation pipeline for any genomics lab. *Bioinformatics (Oxford, England)* **25**, 962-963.

Swain, C. L. & Martin, D. R. (2007). Survival of meningococci outside of the host: implications for acquisition. *Epidemiol Infect* **135**, 315-320.

Swartley, J. S., Ahn, J. H., Liu, L. J., Kahler, C. M. & Stephens, D. S. (1996). Expression of sialic acid and polysialic acid in serogroup B *Neisseria meningitidis*: divergent transcription of biosynthesis and transport operons through a common promoter region. *J Bacteriol* **178**, 4052-4059.

Swartley, J. S., Marfin, A. A., Edupuganti, S., Liu, L. J., Cieslak, P., Perkins, B., Wenger, J. D. & Stephens, D. S. (1997). Capsule switching of *Neisseria meningitidis*. *Proc Natl Acad Sci U S A* **94**, 271-276.

Swartley, J. S., Liu, L. J., Miller, Y. K., Martin, L. E., Edupuganti, S. & Stephens, D. S. (1998). Characterization of the gene cassette required for biosynthesis of the (alpha1-->6)-linked N-acetyl-D-mannosamine-1-phosphate capsule of serogroup A *Neisseria meningitidis*. *J Bacteriol* **180**, 1533-1539.

Takaya, A., Suzuki, M., Matsui, H., Tomoyasu, T., Sashinami, H., Nakane, A. & Yamamoto, T. (2003). Lon, a stress-induced ATP-dependent protease, is critically important for systemic *Salmonella enterica* serovar typhimurium infection of mice. *Infect Immun* **71**, 690-696.

Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**, 1596-1599.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D. & other authors (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41.

Tettelin, H., Saunders, N. J., Heidelberg, J. & other authors (2000). Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* **287**, 1809-1815.

Tettelin, H., Masignani, V., Cieslewicz, M. & other authors (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America* **102**, 13950-13955.

Thompson, E. A., Feavers, I. M. & Maiden, M. C. (2003). Antigenic diversity of meningococcal enterobactin receptor FetA, a vaccine component. *Microbiology* **149**, 1849-1858.

Thompson, J. D., Gibson, T. J. & Higgins, D. G. (2002). Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics* **Chapter 2**, Unit 2.3.1-2.3.22.

Tikhomirov, E., Santamaria, M. & Esteves, K. (1997). Meningococcal disease: public health burden and control. *World Health Stat Q* **50**, 170-177.

Tinsley, C. R., Voulhoux, R., Beretti, J. L., Tommassen, J. & Nassif, X. (2004). Three homologues, including two membrane-bound proteins, of the disulfide oxidoreductase DsbA in *Neisseria meningitidis*: effects on bacterial growth and biogenesis of functional type IV pili. *J Biol Chem* **279**, 27078-27087.

Tsang, R. S., Law, D. K., Tyler, S. D., Stephens, G. S., Bigham, M. & Zollinger, W. D. (2005). Potential capsule switching from serogroup Y to B: The characterization of three such *Neisseria meningitidis* isolates causing invasive meningococcal disease in Canada. *Can J Infect Dis Med Microbiol* **16**, 171-174.

Tsang, R. S., Tsai, C. M., Henderson, A. M., Tyler, S., Law, D. K., Zollinger, W. & Jamieson, F. (2008). Immunochemical studies and genetic background of two *Neisseria meningitidis* isolates expressing unusual capsule polysaccharide antigens with specificities of both serogroup Y and W135. *Can J Microbiol* **54**, 229-234.

Tsilibaris, V., Maenhaut-Michel, G. & Van Melderen, L. (2006). Biological roles of the Lon ATP-dependent protease. *Res Microbiol* **157**, 701-713.

Tzeng, Y. L., Datta, A. K., Strole, C. A., Lobritz, M. A., Carlson, R. W. & Stephens, D. S. (2005). Translocation and surface expression of lipidated serogroup B capsular Polysaccharide in *Neisseria meningitidis*. *Infect Immun* **73**, 1491-1505.

Uniprot, C. (2009). The Universal Protein Resource (UniProt) 2009. *Nucl Acids Res* **37**, D169-174.

Vernikos, G. S. & Parkhill, J. (2006). Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics* **22**, 2196-2203.

Vieusseux, M. (1806). Mémoire sur la maladie qui a regné a Genêve au printemps de 1805. *J Med Chir Pharmacol* **11**, 163-182.

Vivian, J. P., Scoullar, J., Rimmer, K. & other authors (2009). Structure and Function of the Oxidoreductase DsbA1 from *Neisseria meningitidis*. *J Mol Biol* **394**, 931-943.

Vogel, U., Weinberger, A., Frank, R., Muller, A., Kohl, J., Atkinson, J. P. & Frosch, M. (1997). Complement factor C3 deposition and serum resistance in isogenic capsule and lipooligosaccharide sialic acid mutants of serogroup B *Neisseria meningitidis*. *Infect Immun* **65**, 4022-4029.

Vogel, U., Claus, H. & Frosch, M. (2000). Rapid serogroup switching in *Neisseria meningitidis*. *N Engl J Med* **342**, 219-220.

Wang, H., Su, Y., Mackey, A. J., Kraemer, E. T. & Kissinger, J. C. (2006). SynView: a GBrowse-compatible approach to visualizing comparative genome data. *Bioinformatics* **22**, 2308-2309.

Warren, L. & Blacklow, R. S. (1962). The biosynthesis of cytidine 5'-monophospho-n-acetylneuraminic acid by an enzyme from *Neisseria meningitidis*. *J Biol Chem* **237**, 3527-3534.

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191.

Weber, M. V., Claus, H., Maiden, M. C., Frosch, M. & Vogel, U. (2006). Genetic mechanisms for loss of encapsulation in polysialyltransferase-gene-positive meningococci isolated from healthy carriers. *Int J Med Microbiol* **296**, 475-484.

Weichselbaum, A. (1887). Üeber die Aetiologie der akuten *Meningitis cerebrospinalis*. *Fortschr Med* **5**, 573-587.

Wu, H. M., Harcourt, B. H., Hatcher, C. P. & other authors (2009). Emergence of ciprofloxacin-resistant *Neisseria meningitidis* in North America. *N Engl J Med* **360**, 886-892.

Yang, J., Chen, L., Sun, L., Yu, J. & Jin, Q. (2008). VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. *Nucleic Acids Res* **36**, D539-542.

Yazdankhah, S. P., Kriz, P., Tzanakaki, G. & other authors (2004). Distribution of serogroups and genotypes among disease-associated and carried isolates of *Neisseria meningitidis* from the Czech Republic, Greece, and Norway. *J Clin Microbiol* **42**, 5146-5153.

Yu, N. Y., Wagner, J. R., Laird, M. R. & other authors (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **26**, 1608-1615.

Zdobnov, E. M. & Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847-848.

Zerbino, D. R. & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829.

Zimin, A. V., Smith, D. R., Sutton, G. & Yorke, J. A. (2008). Assembly reconciliation. *Bioinformatics* **24**, 42-45.

Zombre, S., Hacen, M. M., Ouango, G., Sanou, S., Adamou, Y., Koumare, B. & Konde, M. K. (2007). The outbreak of meningitis due to *Neisseria meningitidis* W135 in 2003 in Burkina Faso and the national response: main lessons learnt. *Vaccine* **25 Suppl 1**, A69-71.